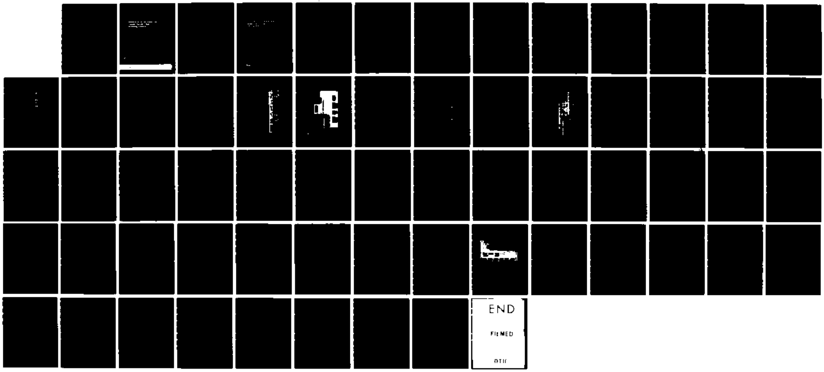
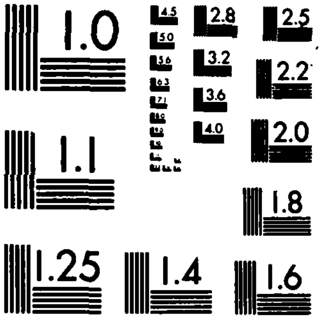
AD-A145 817 INTRODUCTION TO THE SUBJECTIVE TRANSFER FUNCTION APPROACH TO ANALYZING SYSTEMS(U) RAND CORP SANTA MONICA CA C T VEIT ET AL. JUL 84 RAND/R-3021-AF 1/1

UNCLASSIFIED F49620-82-C-0018 F/G 5/1 NL





1.0
4.5
5.0
5.6
6.3
7.1
8.0
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

12

R-3021-AF

Introduction to the Subjective Transfer Function Approach to Analyzing Systems

Clairice T. Veit, Monti Callero, Barbara J. Rose

July 1984

Rand PROJECT AIR FORCE

The research reported here was sponsored by the Directorate of Operational Requirements, Deputy Chief of Staff/Research, Development, and Acquisition, Hq USAF, under Contract F49620-82-C-0018. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

Library of Congress Cataloging in Publication Data

Veit, Clairice T.
Introduction to the subjective transfer function approach
to analyzing systems.

"Prepared for the United States Air Force."
Bibliography: p.
"R-3021-AF."
1. Subjective transfer function method. 2. System
analysis. I. Callero, Monti. II. Rose, Barbara J.
III. Rand Corporation. IV. United States. Air Force.
QA402.V44 1984 003 84-6776
ISBN 0-8330-0553-7

The Rand Publications Series: The Report is the principal publication documenting and transmitting Rand's major research findings and final research results. The Rand Note reports other outputs of sponsored research for general distribution. Publications of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

Published by The Rand Corporation

R-3021-AF

Introduction to the Subjective Transfer Function Approach to Analyzing Systems

Clairice T. Veit, Monti Callero, Barbara J. Rose

July 1984

A Project AIR FORCE report
prepared for the
United States Air Force

Rand
1700 MAIN STREET
P.O. BOX 2138
SANTA MONICA, CA 90406-2138

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER R-3021-AF	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Introduction to the Subjective Transfer Function Approach to Analyzing Systems		5. TYPE OF REPORT & PERIOD COVERED Interim
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Clairice T. Veit, Monti Callero, Barbara J. Rose		8. CONTRACT OR GRANT NUMBER(s) F49620-82-C-0018
9. PERFORMING ORGANIZATION NAME AND ADDRESS The Rand Corporation 1700 Main Street Santa Monica, CA 90406		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Requirements, Programs and Studies Group (AF/RDQM) Ofc, DSC/RED And Acquisition Hq USAF, Washington, DC 20330		12. REPORT DATE July 1984
		13. NUMBER OF PAGES 54
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release; Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

No Restrictions

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Systems Analysis
Systems Approach
Judgment

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

see reverse side

The subjective transfer function (STF) approach is a general subjective measurement method developed to evaluate complex systems where groups of people differing in expertise directly and indirectly affect system outcomes. The approach is based on the pinciples of hypothesis formulation and testing. It incorporates features of the algebraic modeling approach to measurement where meaningful subjective scale values derive from tested theories; it provides additional features necessary for coalescing judgments obtained from different groups of system experts into an overall perceptual outcome. This report introduces and is a primer of the STF method. It outlines the steps involved in the approach, describes how those steps can be accomplished, and discusses measurement principles and techniques to aid the reader's understanding of the basis for the approach.

PREFACE

The Subjective Transfer Function (STF) approach to analyzing systems was developed at Rand during research on evaluating the contribution of command and control to the overall combat effectiveness of tactical air forces. This report is intended to serve as a primer of the STF method for personnel faced with the problem of analyzing systems wherein hard (equipment), soft (procedures), and human elements all have main effects on outcomes. The work was sponsored by Project AIR FORCE under the study effort "Tactical Air Command and Control."

SUMMARY

The subjective transfer function (STF) approach is a general subjective measurement method developed to evaluate complex systems where groups of people differing in expertise directly and indirectly affect system outcomes. The approach is based on the principles of hypothesis formulation and testing. It incorporates features of the algebraic modeling approach to measurement where meaningful **subjective scale values derive from tested theories**; it provides additional features necessary for coalescing judgments obtained from different groups of system experts into an overall perceptual outcome.

In the STF approach, a complex system is divided into units corresponding to tasks performed by different groups of system "experts." Factors and outcomes describing each unit are identified with a body of experts; causal hypotheses in the form of algebraic functions that specify effects of those factors on judged outcomes are *tested* and rejected if they are not supported by the data. The algebraic functions that describe the interrelationships between factors and outcomes in each unit are used to predict outcomes under different conditions described by different system capabilities. They provide a basis for assessing extant capabilities, investigating the effects of system and concept modifications, and determining configuration alternatives or function requirements to meet system objectives.

This report introduces and is a primer of the STF method. It outlines the steps involved in the approach, describes how those steps can be accomplished, and discusses measurement principles and techniques to aid the reader's understanding of the basis for the approach. The report is organized as follows:

- **Development of initial and alternative structural and functional (subjective transfer function) system hypotheses.**
- **Examples of how to design judgment experiments so as to provide tests of those hypotheses.**
- **Discussion on determining STFs and the final system structure.**
- **How final structure and functions are used to evaluate a system.**

The last section briefly discusses two other subjective measurement approaches commonly used for analyzing systems, their measurement problems, and how the STF approach provides resolutions to those problems.

A1

CONTENTS

PREFACE iii

SUMMARY v

FIGURES AND TABLES ix

Section

I. OVERVIEW OF THE SUBJECTIVE TRANSFER APPROACH 1

II. DEVELOPING STRUCTURAL AND FUNCTIONAL HYPOTHESES 3
Structural Hypotheses 3
STF Hypotheses 12

III. CONDUCTING JUDGMENT EXPERIMENTS 17
Designing Experiments to Test Hypotheses 17
Selecting Experimental Designs *31*
Collecting Judgment Data 33

IV. DETERMINING STFs AND FINAL STRUCTURE 34
Data Analyses 34
Example of a Final System Structure 38

V. EVALUATING SYSTEMS 41

VI. COMMENTS ON OTHER SUBJECTIVE APPROACHES TO ANALYZING . . . .
SYSTEMS 50
Using "Direct" Scales and Assumed Models 50
Saaty's Approach 50
Summary Remarks 51

BIBLIOGRAPHY 53

FIGURES

1. Hypothesized Immediate Targeting Structure 4
2. Structure Depicting Hypothesized Factors Affecting the Likelihood of Joining the Air Force 7
3. An Alternative Initial Structure to Fig. 2 8
4. An Alternative Initial Structure to Figs. 2 and 3 9
5. Alternative Structure to Fig. 1 11
6. Alternative Structure to Fig. 3 13
7. Final Structure of a Command and Control System 14
8. Outline of Subjective Measurement 15
9. Example of Factorial Designs 18
10. Hypothetical Data that Might Be Obtained from 2-Way Factorial Designs 20
11. Experimental Design that Varies Amount of Information 22
12. Predictions of Four Different Functions When Designs Vary the Amount of Information 24
13. Graphic Predictions of Additive and Averaging Functions 26
14. Scale-Free Design for the Enemy Emitter Location and Coverage Factors 29
15. Hypothetical Data Predicted by Different Algebraic Functions 35
16. Alternative Structure/Function Combinations 37
17. Graphical Assessment of the Range Model 39
18. Final Structure and STRs Corresponding to Fig. 1 40
19. System Comparisons 42
20. Subjective Scale Values for the Location/Classification, Coverage, and Currency Factors Shown in Fig. 1 45
21. Theoretical Predictions 46
22. One Section of the Command and Control System Shown in Fig. 7 48
23. Same Section of Fig. 7 as Shown in Fig. 22 49

TABLES

1. Definitions of Factors and Outcomes for Immediate Targeting Structure Shown in Figs. 1 and 3 5
2. Possible STFs 16

I. OVERVIEW OF THE SUBJECTIVE TRANSFER FUNCTION APPROACH

The Subjective Transfer Function (STF) approach is a subjective measurement method that relies on "expert" judgment for analyzing complex systems where many factors either directly or indirectly affect system outcomes. Some examples of complex systems that have received attention in the literature are military command and control, education, transportation, and management. Examples of evaluation interests would be to determine how such Air Force command and control factors as communication and information systems affect the performance of air missions and how such factors of a management system as computer processing capabilities and employee training programs affect job performance.

A major problem in any evaluation procedure is to define the system in terms of its factors and outcomes in such a way that causal hypotheses about effects of those factors on the outcomes can be *tested* and rejected if the data do not support them. This feature of hypothesis testing is an integral part of the STF approach, which is based on the algebraic modeling approach to measurement. The idea behind this approach is that subjective scale values associated with information being judged have meaning with respect to the theory that describes the judgment process. When a judgment theory passes stringent tests of its predictions, the scale values are a by-product of the theory and have substantive meaning with respect to the theory. This approach includes functional measurement (Anderson, 1970, 1979, 1981), conjoint measurement (Krantz et al., 1971; Krantz and Tversky, 1971), and the principles of stimulus scale convergence and scale-free tests (Birnbaum, 1974; Birnbaum and Veit, 1974a,b), which are important tools for testing judgment theories.

In the STF approach, complex systems are analyzed from the perspective of the "expert," who by definition knows and understands the system. Typically, different groups of experts know about different aspects of the system. Experts from each group make judgments about outcomes resulting from their tasks that would be expected under different descriptions of system capabilities. The judgment theory (STF) for each expert group specifies the effects of the different system capabilities on these judged outcomes. The set of STFs across expert groups links the outcomes associated with different tasks within the system to an outcome(s) corresponding to a measure(s) of overall system effectiveness.

The steps involved in the STF approach can be outlined as follows.

- Develop an initial structure of the system. This requires identifying the system effectiveness outcomes of interest and postulating the factors thought to affect them.
- Postulate hypotheses about *how* factors affect the expert's perception of those outcomes. These are in the form of algebraic functions (referred to as subjective transfer functions) that specify how subjective values the "expert" associates with the factors are combined to form a perception about the outcome.
- Construct experimental designs that permit *tests* among the alternative hypotheses.
- Collect judgment data. The experimental designs are incorporated into a paper and pencil questionnaire format and fielded to the "expert" respondents.
- Analyze judgment data to determine the complete model (the system factors and the STFs that link factors to outcomes) of the system that best explains the data.

- **Evaluate system capabilities using the model.** Once a complete model of the system has been determined, it is used to evaluate how different system capabilities (defined by different factor descriptions) affect the outcomes, and their tradeoffs in affecting those outcomes.

II. DEVELOPING STRUCTURAL AND FUNCTIONAL HYPOTHESES

A complex system structure depicts the factors that make up the system and the direct and indirect effects they have on system outcomes. The STFs specify the causal links among the factors and outcomes. Alternative structures and STFs have to be hypothesized before data are gathered, because tests of the hypotheses are possible only when experiments have been appropriately designed. In this section, we discuss procedures involved in developing hypotheses of system structures and STFs.

STRUCTURAL HYPOTHESES

Structural development requires interaction with system experts. For many complex systems different people are expert in different parts of the system; we therefore develop a structure of each part of the system in conjunction with a body of experts. The researcher's job is to work with the experts to identify and define important system factors and outcomes that both make sense to the expert and can be manipulated in experimental designs.

Identifying Outcomes and Factors

The first step is to identify the outcomes produced by the system that provide the important external measures of the system's effectiveness. Next one identifies factors thought to directly affect these outcomes. Some or all of these factors may represent outcomes that are produced within the system (referred to as *suboutcomes)* and are themselves affected by other system factors. A hierarchical causal representation of the system develops when system factors are identified for suboutcomes until all suboutcomes are affected only by factors that represent system input characteristics or basic system features.[1] Such factors are called *primitive factors.*

As an example, Fig. 1 shows a structure for a tactical air command and control process that was investigated using the STF approach (Veit, Callero, and Rose, 1982). The structure contains one factor/outcome set (upper portion of the figure) and one factor/suboutcome set (lower portion of the figure). Such sets are referred to as *experimental units.* In Fig. 1 the two experimental units correspond to two different groups of experts. The group corresponding to experimental unit 1 (upper portion) performs the immediate targeting task of pairing tactical aircraft with important enemy ground force targets in a timely manner.[2] The single overall measure of system effectiveness in this example is how well U.S. Air Force officers perceive they can perform their immediate targeting task under various conditions having to do with the information and equipment they work with. The experts corresponding to experimental unit 2 (darker section—Target Identification) are expert in identifying enemy targets.

[1] If there are system inputs or basic system features that are of particular interest with respect to their effects on specified system outcomes (for example, characteristics that correspond to equipment being considered for purchase), they need to be included in whatever detail is necessary to satisfy the evaluation goals.

[2] Immediate targeting involves recognizing that an important target exists, determining the availability of tactical aircraft that have the proper weapons to destroy the target in the prevailing weather conditions, and directing the aircraft to attack the target.

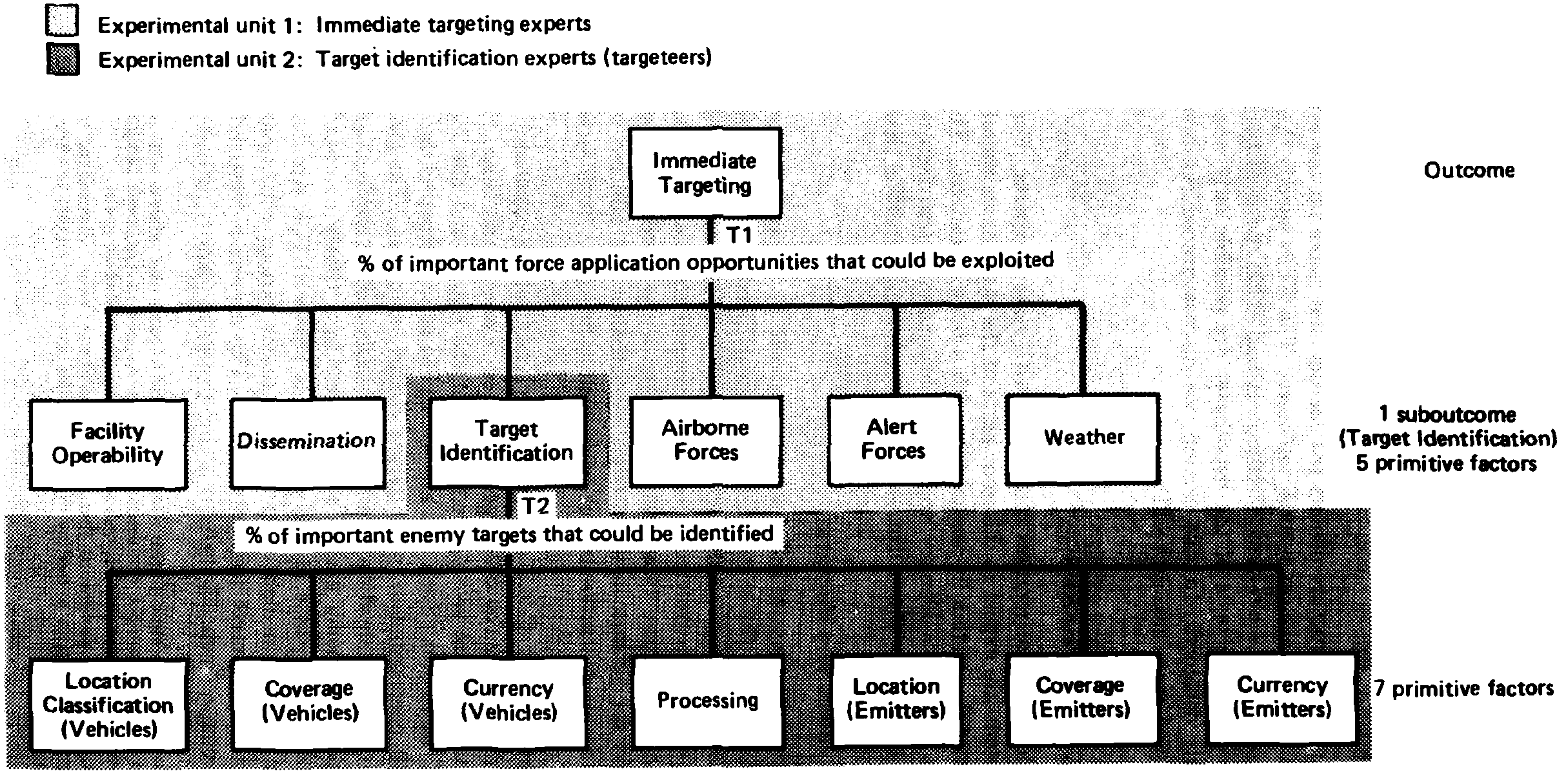

□ Experimental unit 1: Immediate targeting experts
▨ Experimental unit 2: Target identification experts (targeteers)
Immediate Targeting
T1
% of important force application opportunities that could be exploited
Outcome
Facility Operability
Dissemination
Target Identification
Airborne Forces
Alert Forces
Weather
1 suboutcome (Target Identification) 5 primitive factors
T2
% of important enemy targets that could be identified
Location Classification (Vehicles)
Coverage (Vehicles)
Currency (Vehicles)
Processing
Location (Emitters)
Coverage (Emitters)
Currency (Emitters)
7 primitive factors

Fig. 1—Hypothesized immediate targeting structure

In experimental unit 1, six factors are hypothesized to affect Immediate Targeting directly; in experimental unit 2, seven factors are hypothesized to affect Immediate Targeting indirectly, through the suboutcome Target Identification.[3] Their definitions are presented in Table 1. (Outcomes define the judgment task; factors define system capabilities.)

Because factors are manipulated in experimental designs (described and illustrated later), levels have to be identified for each factor. Factor levels that were selected for the factors shown in Fig. 1 are presented next to each factor definition in Table 1. The factor levels should span the range from the worst to the best capability that might be expected over the time period of interest. This feature is important if future conditions or characteristics of future systems are to be built into the model for evaluation purposes. Selection of factor levels between the endpoints may be guided by such things as descriptions of equipment actually being considered for research, production, or purchase, and descriptions of existing equipment capabilities. From three to five factor levels are usually sufficient for experimental purposes, as will be discussed in the section on experimental design.

Table 1

DEFINITIONS OF FACTORS AND OUTCOMES FOR IMMEDIATE TARGETING STRUCTURE SHOWN IN FIGS. 1 AND 3

A. Experimental Unit 1 (Immediate Targeting Experts)

Judged Outcome: The percent of force application opportunities that could be exploited in a timely manner

Factor Definitions	Factor Levels
Target Identification (percent of important force elements identified)	90 60 30 10
Facility Operability (percent of immediate targeting activities that can be supported by the facility)	90 60 30 10
Alert Forces (status of the Alert Forces accessible in the C^2 facility)	90 60 30 10
Airborne Forces (status of the airborne forces accessible in the C^2 facility)	90 60 30 10
Weather (currency of the reliable weather information)	15 min., 1 hr., 3 hrs., 12 hrs.
Dissemination (percent of the forces that can be tasked in a timely manner)	90 60 30 10

[3] Factors hypothesized to affect an outcome or suboutcome pertaining to a particular expert respondent population must not exceed the span of knowledge of the particular expert group.

Table 1 (Continued)

B. Experimental Unit 2 (Targeting Experts)

Judged Outcome: The percent of important enemy targets that could be identified in a timely manner

Factor Definitions	Factor Levels
Vehicle Location/Classification (ability of sensor systems to locate and classify enemy vehicles)	Locate and classify in all weather Locate (not classify) in all weather Locate and classify in clear weather
Vehicle Coverage (percent of enemy vehicles that have been observed)	90 60 30 10
Vehicle Currency (time interval between the observation of enemy vehicles and the data's availability for processing)	5 min., 15 min., 30 min., 1 hr.
Processing (the means by which enemy vehicle and emitter information is interpreted)	Fully computerized interpretation. Human uses computer to graphically display information; human interpretation. Human uses computer to sort textual information; human interpretation. Human sorts hard copy, textual information; human interpretation.
Emitter Location Accuracy (accuracy with which enemy emitters are located)	10m, 100m, 1000m
Emitter Coverage (percent of the enemy emitters that have been observed)	90 60 30 10
Emitter Currency (time interval between the observation of emitters and the data's availability for processing)	5 min., 15 min., 30 min., 1 hr.

The structural hypothesis shown in Fig. 1 says that part of the ability to do immediate targeting depends on how well the targeteers are able to identify important enemy targets. Target Identification is a contributing factor in the first experimental unit and an outcome in the second experimental unit. When a factor serves this dual purpose, it is necessary to define it in the *same* terms (see Table 1) for the two groups of experts to satisfy the transfer feature of the STFs (discussed later).

The structural representation shown in Fig. 1 hypothesizes *two* STFs. The first (T1) specifies the causal link among the six factors affecting the immediate targeting outcome, and the second (T2) specifies the causal link among the seven factors affecting Target Identification. These functions are referred to as *subjective* because they are models of judgment processes that are not directly observed occurring between the time a stimulus is perceived (e.g., a description of the system's capabilities) and the time a response occurs (e.g., judgment of task performance). They are referred to as *transfer* functions because, when their functional forms have been determined and they are being computed to evaluate a particular system, the output of one function transfers for use as an input value to the function above it. For example, the output of T2 in Fig. 1 would identify the target identification factor level needed to determine that factor's subjective input value to T1. (Examples of using STFs are presented later.)

Figures 2–4 illustrate a problem commonly encountered when one is structuring complex systems. At issue in these problem domains is the likelihood an individual will join the Air Force under varied benefit packages. The factors selected would reflect benefits being

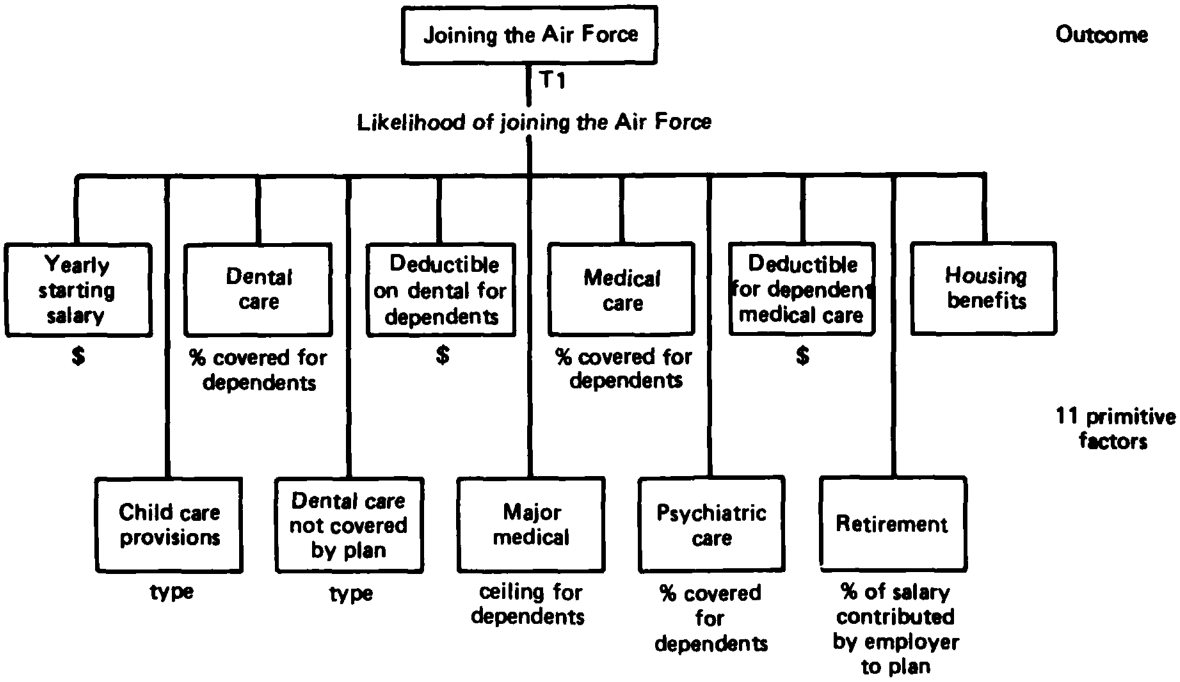


Fig. 2—Structure depicting hypothesized factors affecting the likelihood of joining the Air Force

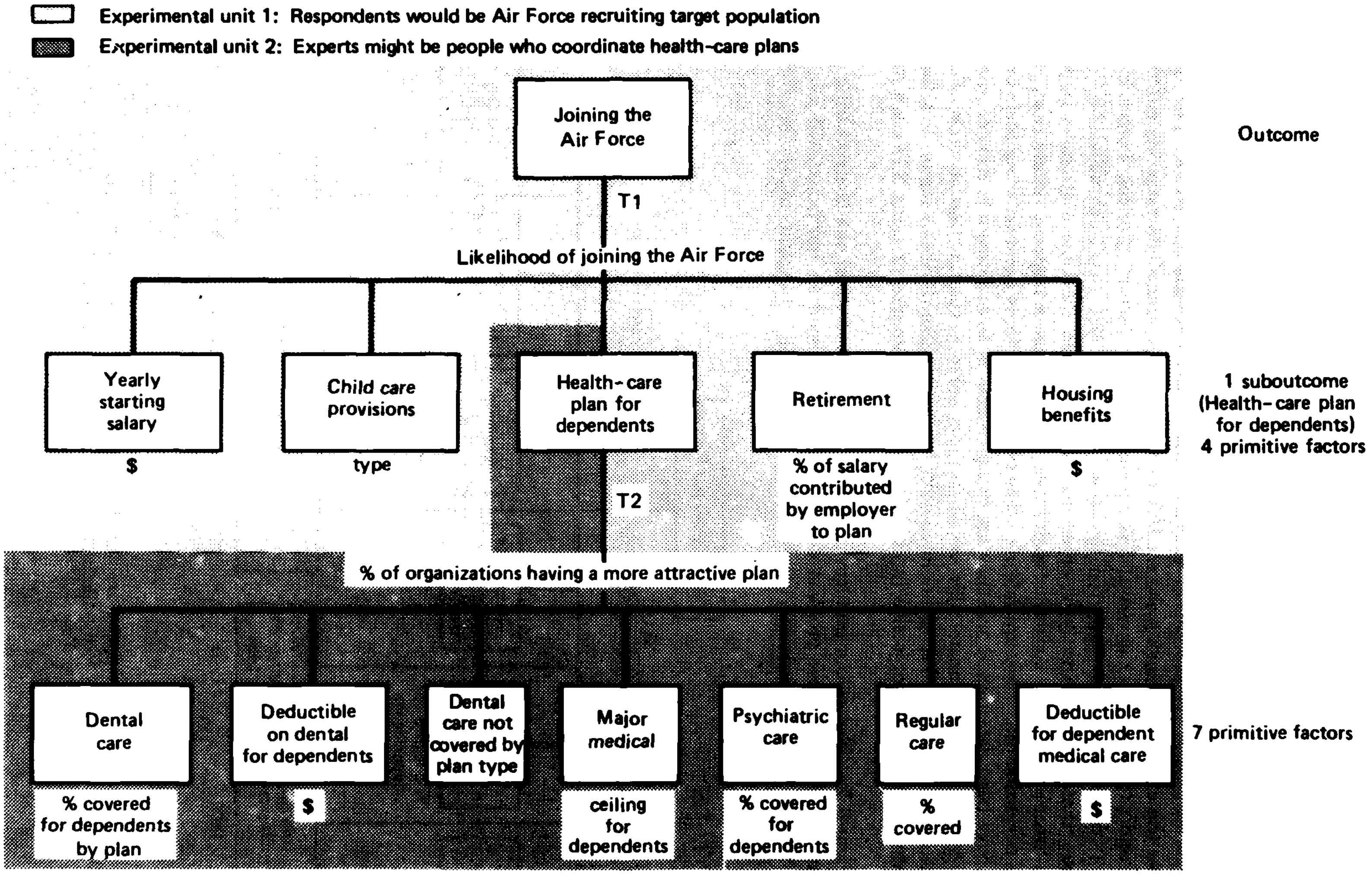
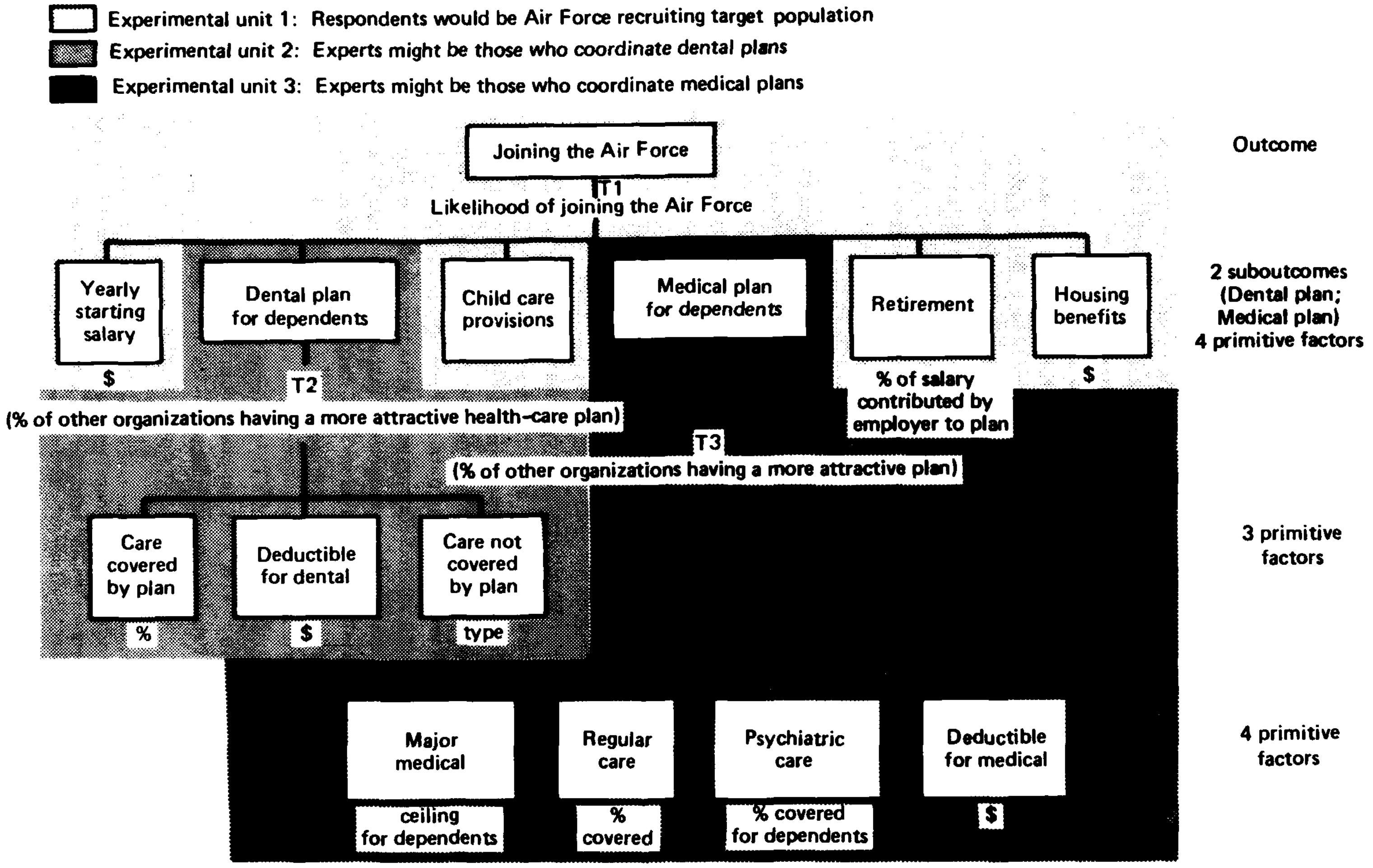

Experimental unit 1: Respondents would be Air Force recruiting target population
Experimental unit 2: Experts might be people who coordinate health-care plans
Joining the Air Force
Outcome
T1
Likelihood of joining the Air Force
Yearly starting salary
$
Child care provisions
type
Health-care plan for dependents
Retirement
% of salary contributed by employer to plan
Housing benefits
$
1 suboutcome (Health-care plan for dependents) 4 primitive factors
T2
% of organizations having a more attractive plan
Dental care
% covered for dependents by plan
Deductible on dental for dependents
$
Dental care not covered by plan type
Major medical
ceiling for dependents
Psychiatric care
% covered for dependents
Regular care
% covered
Deductible for dependent medical care
$
7 primitive factors

Fig. 3—An alternative initial structure to Fig. 2
(depicts two experimental units)



□ Experimental unit 1: Respondents would be Air Force recruiting target population
▨ Experimental unit 2: Experts might be those who coordinate dental plans
■ Experimental unit 3: Experts might be those who coordinate medical plans
Joining the Air Force
Outcome
T1
Likelihood of joining the Air Force
Yearly starting salary
$
Dental plan for dependents
T2
Child care provisions
Medical plan for dependents
Retirement
% of salary contributed by employer to plan
Housing benefits
$
2 suboutcomes (Dental plan; Medical plan) 4 primitive factors
(% of other organizations having a more attractive health-care plan)
T3
(% of other organizations having a more attractive plan)
Care covered by plan
%
Deductible for dental
$
Care not covered by plan
type
3 primitive factors
Major medical
ceiling for dependents
Regular care
% covered
Psychiatric care
% covered for dependents
Deductible for medical
$
4 primitive factors

Fig. 4—An alternative initial structure to Figs. 2 and 3
(depicts three experimental units)

considered for modification, say by the Congress. The idea is to *measure* the effects of the factors on judged likelihood of joining the Air Force.[4] This structure depicts 11 factors affecting the judged outcome. Experimental manipulation of the factors requires that some items presented to respondents for judgment contain a factor level of each factor. Eleven job characteristics seem to be too many to judge simultaneously,[5] making it necessary to reduce the number of factors considered at one time.

We can reduce the number of factors by hypothesizing a subset of them to affect one or more suboutcomes that are meaningful within the framework of the evaluation goals. For example, the seven factors describing the health-care plan could be hypothesized to affect perceptions of the percent of competing organizations that have a better plan, thus creating the initial structure shown in Fig. 3. The second experimental unit in Fig. 3 has a different respondent population associated with it—people who coordinate health-care plans.

The dependent variable defined for the new experimental unit should be meaningful for the group that will judge the factor as a dependent variable (e.g., health-plan coordinators for experimental unit 2 in Fig. 3) and for the group that will judge it as an independent variable (Air Force recruiting target population). The creation of a new experimental unit need not result in a new respondent population. When the same respondent group is expert in both outcomes, they could be asked to make judgments about their two different tasks at different times, or different subsets of the expert group could be assigned to the different experimental units.

Another alternative would be to separate the health-care plan suboutcome into a dental plan and a medical plan suboutcome, if it is of interest to investigate how each plan individually contributes to the decision to join the Air Force. These two types of benefits could be depicted in separate factor/suboutcome sets, as in Fig. 4. The structure shown in Fig. 4 is the most appealing initial structure of this problem domain because it contains the fewest factors associated with any one respondent group (experimental unit).

Identifying Alternative Structures

Structural alternatives refer to alternative hypotheses about the number of STFs linking factors to outcomes. Alternative structural hypotheses arise from different hypotheses about how the expert combines information included in a description of a system's capabilities. Some of these hypotheses emerge through interaction with the expert respondents during structure development. Others emerge during data analysis, illustrated later.

Consider as an example the immediate targeting structure shown in Fig. 1. An alternative structural hypothesis for the Air Force targeteers (experimental unit 2) is depicted in the lower portion of Fig. 5. This structure proposes that targeteers combine information about enemy emitters such as radars and radios separately from information about enemy vehicles; then they take the subjective values of those outputs and combine them with their value associated with the processing capability factor. This alternative structural hypothesis requires three STFs (T3, T4, and T5). The two separate combination processes are represented diagrammatically by inserting two *intermediary* factors—Vehicles and Emitters—into the structure. (An intermediary factor is one that is *not* identified by factor levels because it is not

[4] Such factors as active members' medical care that are held constant (not manipulated in experimental designs) are not included in the structure but are presented as background information to set the context for judgment.

[5] Our research has indicated that between five and seven pieces of information (depending on the interrelationships among the factors) are maximum; Miller (1956) has estimated seven ± two pieces of information to be maximum for processing information.

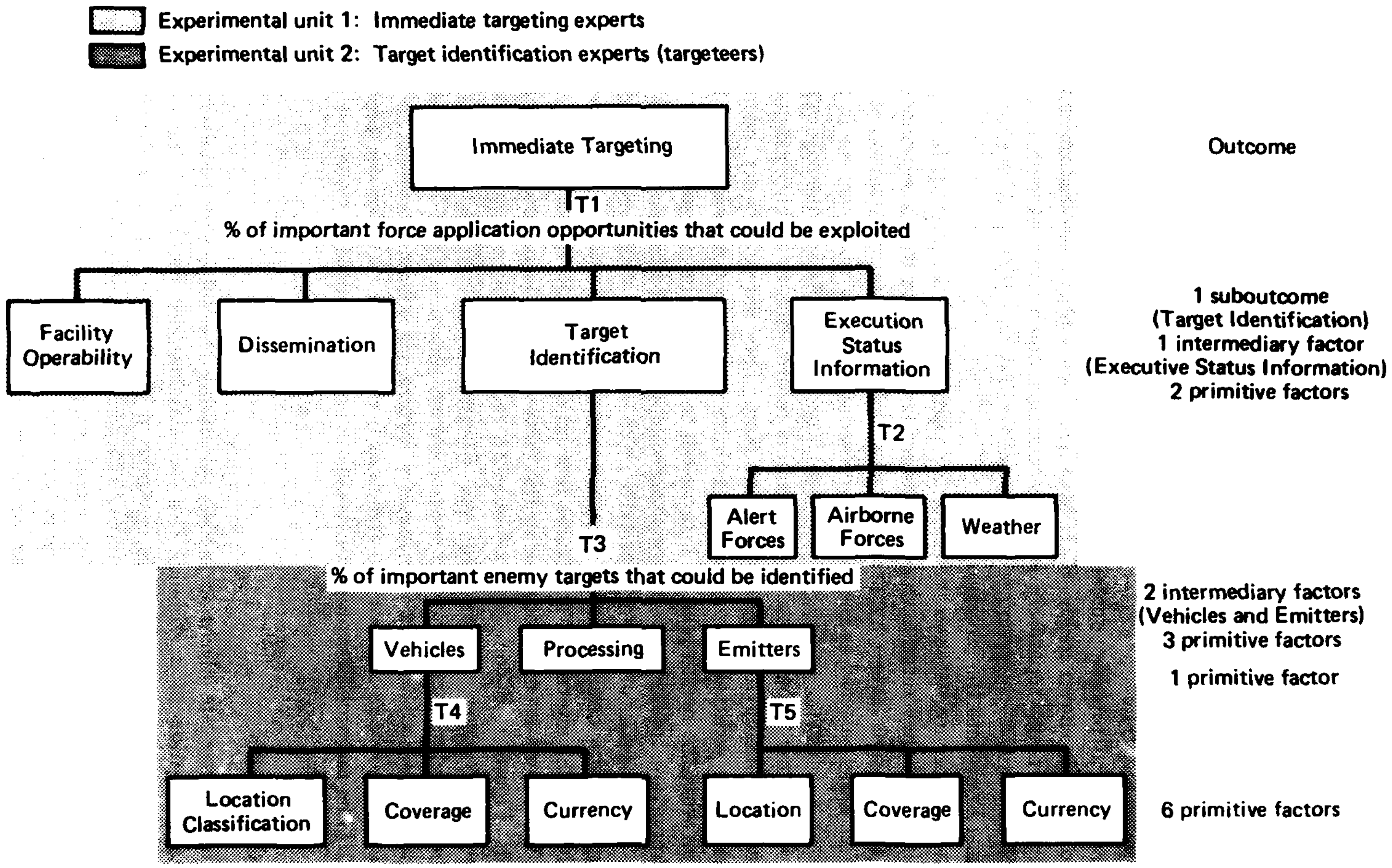

Experimental unit 1: Immediate targeting experts
Experimental unit 2: Target identification experts (targeteers)
Immediate Targeting
Outcome
T1
% of important force application opportunities that could be exploited
Facility Operability
Dissemination
Target Identification
Execution Status Information
1 suboutcome (Target Identification) 1 intermediary factor (Executive Status Information) 2 primitive factors
T2
Alert Forces
Airborne Forces
Weather
T3
% of important enemy targets that could be identified
2 intermediary factors (Vehicles and Emitters) 3 primitive factors
Vehicles
Processing
Emitters
1 primitive factor
T4
T5
Location Classification
Coverage
Currency
Location
Coverage
Currency
6 primitive factors

Fig. 5—Alternative structure to Fig. 1
(depicts three intermediary factors)

manipulated in experimental designs. It depicts the process of separately combining a subset of factors in the experimental unit.)

Other structural hypotheses would change the number of paths and STFs. For example, an alternative structure for experimental unit 1 might result from the hypothesis that an immediate targeting expert (experimental unit 1 of Fig. 1) combines information about the three factors concerning their friendly forces (Alert Forces, Airborne Forces, and Weather) separately from information about their other capabilities—Target Identification, Facility Operability, and Dissemination. This is depicted by inserting the intermediary factor Execution Status Information in this part of the structure. This alternative structure requires two STFs (T1 and T2) for experimental unit 1. Other structural hypotheses would change the number of paths and STFs.

For our second problem domain concerned with Air Force recruitment, an alternative structure to Fig. 3 is shown in Fig. 6. It says health-care plan coordinators combine information about dental plans separately from information about medical plans. This alternative hypothesis is depicted by inserting two intermediary factors (Quality of Dental Plan and Quality of Medical Plan) into the structure.

Note the difference between including the medical- and dental-plan factors in Fig. 4 and in Fig. 6. In Fig. 4 they serve both as *dependent* factors about which respondents make judgments and *independent* factors hypothesized to affect judgments of likelihood of joining the Air Force. If Fig. 4 was selected as the initial representation, they would appear in the final structure as shown. However, in Fig. 6, they are used only to depict one alternative structural hypothesis about how respondents might combine information in making judgments of percent of other organizations having a more attractive plan. They are neither dependent nor independent factors in the judgment experiments and may or may not appear in the final structure, depending on what the health-plan STF turns out to be after the judgment data are analyzed.

It is important to specify as many of these alternative combinatorial hypotheses as seem appropriate before data collection, especially in the larger (4–6 factor) experimental units, so that the experimental designs insure adequate tests among them—that is, provide a basis for knowing which of the alternative structures best accounts for the data.

The illustrations used for discussing the formulation of structural hypotheses have been fairly small. Figure 1 is composed of only 13 factors to be manipulated in experimental designs, one suboutcome, and one outcome. Many systems are much larger than this. For example, the tactical air command and control system shown in Fig. 7 is composed of 12 STFs, 25 factors to be manipulated, six intermediary factors, six suboutcomes, and a final outcome; and that depicts only one-third of the structure hypothesized to affect the final outcome, the land battle. (The research pertaining to this structure is presented in Callero et al., 1984.)

STF HYPOTHESES

STF hypotheses are algebraic models specifying the subjective (unobserved) processes between the perception of a stimulus (e.g., a questionnaire item) and the occurrence of a response. An outline of these processes is shown in Fig. 8. The outline is for three factors but could be extended to include any number. The observed stimuli on the left would be factor levels from three different factors. The outline suggests that the respondent first transforms each factor level (S_i, S_j, S_k) to a scale value (s_i, s_j, s_k) using some function, H (referred to as the utility or psychophysical function), then combines these values according to some combination function (T) to form an integrated impression (r_{ijk}); then the respondent transforms the

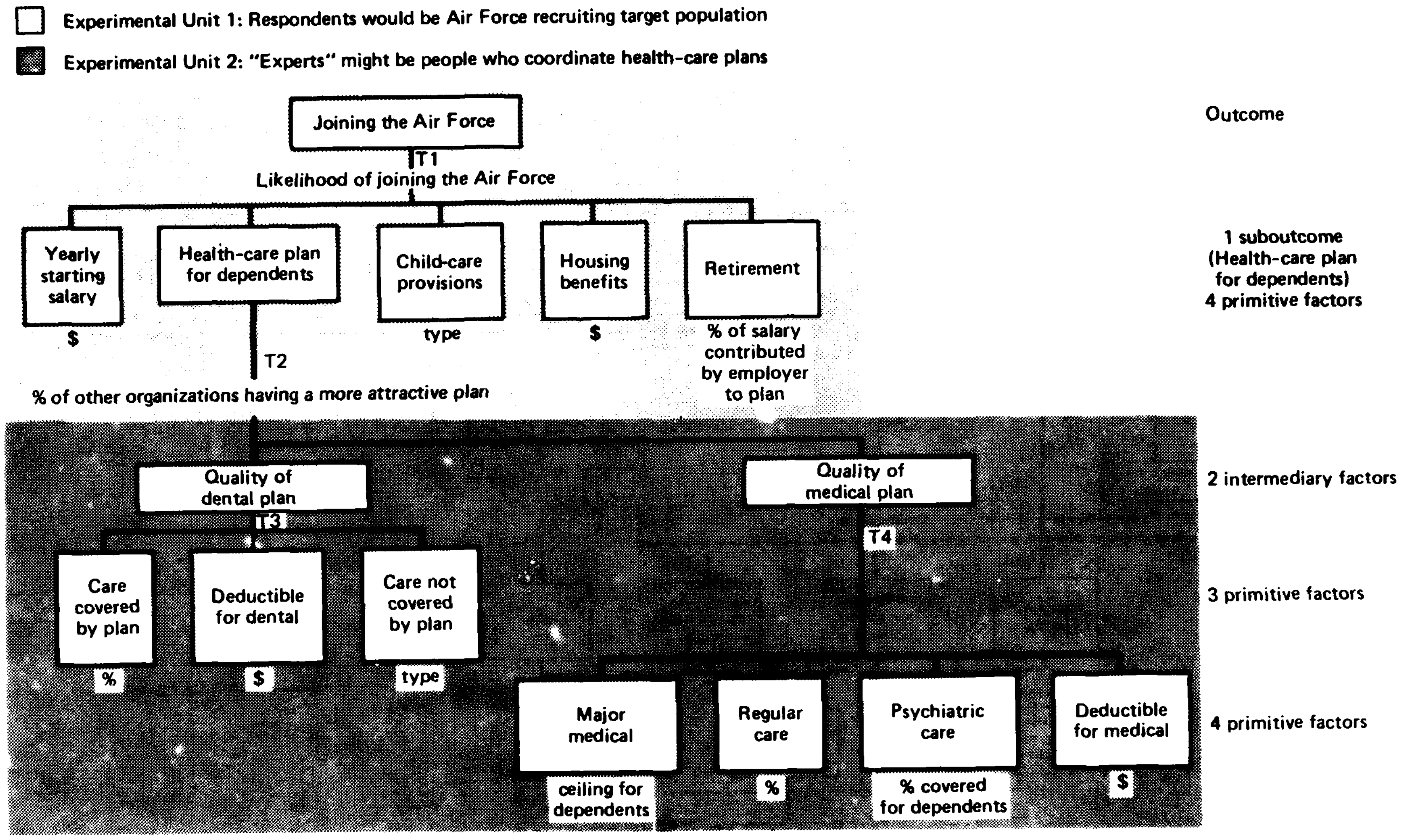


Fig. 6—Alternative structure to Fig. 3
(hypothesizes a different perceptual organization of the factors in experimental unit 2, depicted by the inclusion of two intermediary factors)

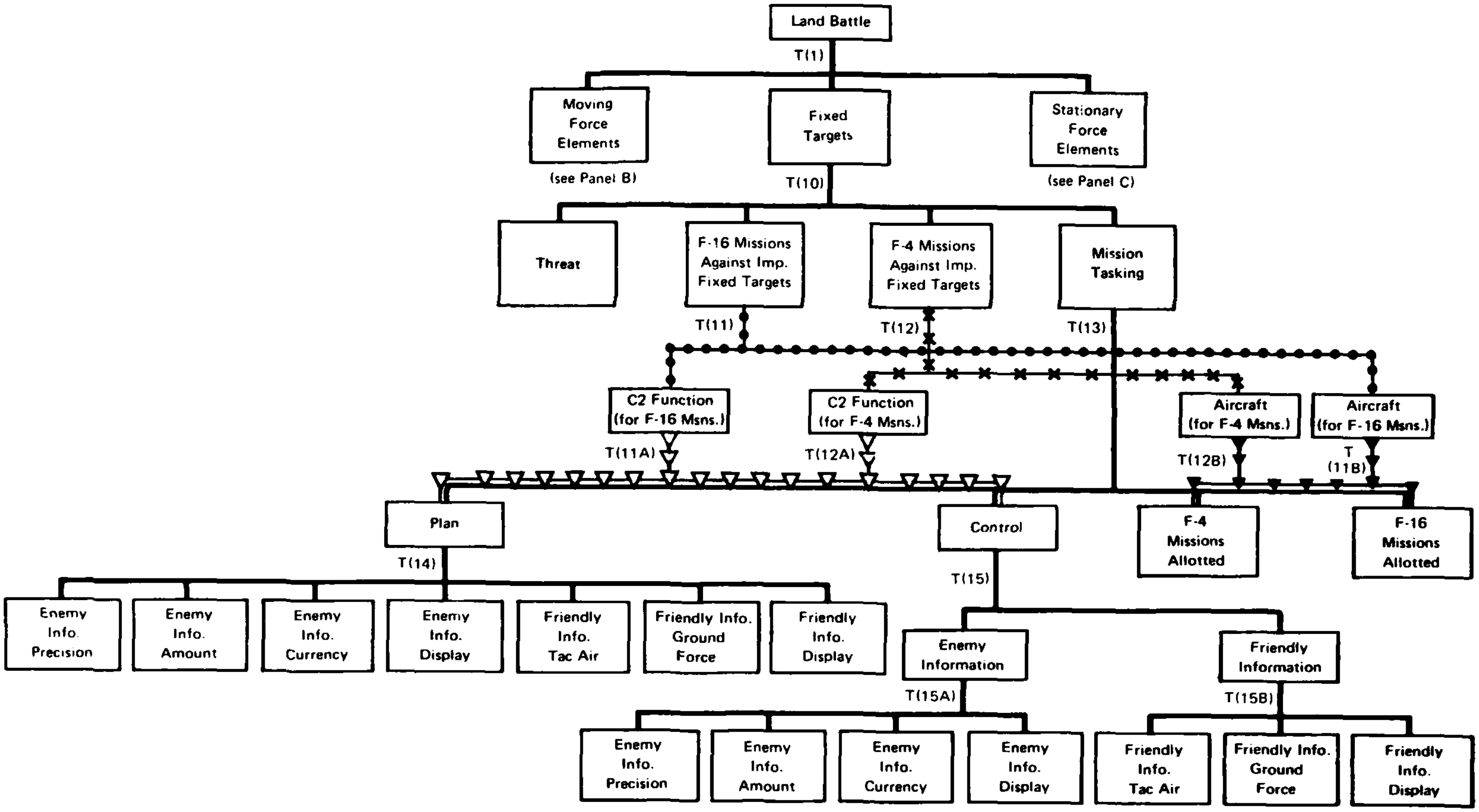


Fig. 7—Final structure of a command and control system
(from Callero et al., 1984)

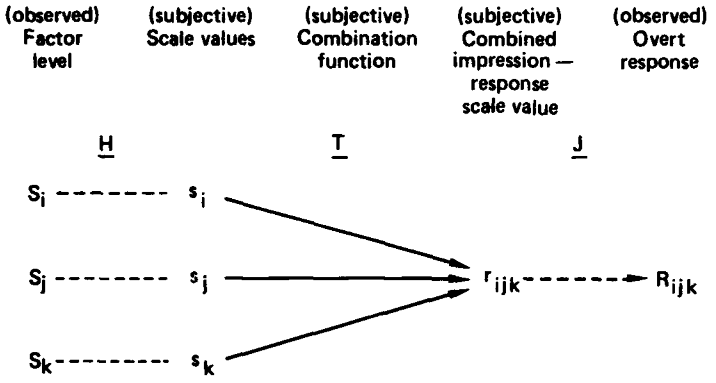

(observed) Factor level
(subjective) Scale values
(subjective) Combination function
(subjective) Combined impression — response scale value
(observed) Overt response
H
T
J
S_i
s_i
S_j
s_j
S_k
s_k
r_{ijk}
R_{ijk}

Fig. 8—Outline of subjective measurement

psychological impression to an overt response (R_{ijk}) by the function, J. The judgment stages can be written:

$$s_i = H(S_i) \quad \text{for stimulus i ,} \tag{1}$$

$$r_{ijk} = T(s_i, s_j, s_k) \ , \tag{2}$$

and

$$R_{ijk} = J(r_{ijk}) \ . \tag{3}$$

Thus, an STF (T in Fig. 8 and Eq. (2)) is a perceptual theory of the expert's judgment process. It specifies how the subjective values the expert places on different factors affect his judgments (e.g., ability to perform a task).

Table 2 describes some algebraic functions that might be considered as STFs at the outset of a complex system investigation. Ideas about what functions to entertain come from the judgment literature and previous research in the problem domain of interest. The functions in Table 2 have been specified for three factors but could be extended to include any number. The r, w, and s parameters are as defined in Eqs. (1)-(3); w_0 and s_0 are "initial estimate" parameters—what the response would be in the absence of specific information. The J function shown in Eq. (1) relating subjective responses, r, to observed responses is not indicated in these equations. Its determination is discussed in the section on the scale-free design.

Each function described in Table 2 makes a *different* prediction with respect to the pattern the judgment data should follow when appropriate experimental designs are used. For example, one prediction all of the functions shown in the upper panel have in common is that of no interactions among the factors. Conversely, the functions in the lower panel can account for interactions among the factors. The functions within each panel make other differential predictions with respect to the judgment data. Some of these are illustrated in the next section.

The form of the STFs considered as possible causal explanations of factor effects on judged outcomes must be specified in advance so that experimental designs allow adequate

Table 2

POSSIBLE STFs

A. Noninteractive functions	
$r = s_{A_j} + s_{B_j} + s_{C_k} + a$	Additive
$r = \dfrac{w_A s_{A_j} + w_B s_{B_j} + w_C s_{C_k}}{w_A + w_B + w_C}$	Averaging
$r = \dfrac{w_0 s_0 + w_A s_{A_j} + w_B s_{B_j} + w_C s_{C_k}}{w_0 + w_A + w_B + w_C}$	Relative-weight (averaging with initial impression)
B. Interactive functions	
$r = s_{A_i} s_{B_j} s_{C_k} + a$	Multiplicative
$r = \dfrac{w_0 s_0 + w_A s_{A_i} + w_B s_{B_j} + w_C s_{C_k}}{w_0 + w_A + w_B + w_C} + \omega(s_{MAX} - s_{MIN})$	Range
$r = \dfrac{w_0 s_0 + w_{A_i} s_{A_i} + w_{B_j} s_{B_j} + w_{C_k} s_{C_k}}{w_0 + w_{A_i} + w_{B_j} + w_{C_k}}$	Differential-weight

*All equations are for three factors: A, B, and C. The parameters of the functions are as follows: s_{A_i}, s_{B_j}, and s_{C_k} are the subjective values for the ith, jth, and kth levels of factors A, B, and C, respectively; w_A, w_B, and w_C, are the weights associated with factors A, B, and C, respectively (a subscript is added when the scale value varies with the factor level); r is the subjective response; w_0 and s_0 are the weight and scale value associated with the initial impression (what the response would be in the absence of specific information); ω denotes the weight of the range term; and a is an additive constant.

tests among their predictions. When tests of an STF support it as an appropriate explanation of respondents' judgments, subjective values associated with the factors and outcomes (the s, w, and r parameters specified in the functions shown in Table 2 and outlined in Fig. 8) are known; they are the least-squares estimates of the function.[6]

[6] Least-squares estimates are preferred over maximum likelihood procedures for estimating parameters primarily because no assumptions are needed on the distribution of the responses. When distributional assumptions are required, as in maximum likelihood, the judgment theory must incorporate those assumptions, resulting in a more complicated theory.

III. CONDUCTING JUDGMENT EXPERIMENTS

After the formulation of the alternative structural and functional hypotheses, experimental designs have to be carefully selected so as to permit tests among the hypotheses. These designs are translated into a paper and pencil questionnaire and fielded to the appropriate expert respondents.

DESIGNING EXPERIMENTS TO TEST HYPOTHESES

The experimental design is crucial to testing among the unique predictions of the STFs under consideration. The design is primarily guided by what the researcher knows about the unique predictions of the STFs. Alternative structural hypotheses guide the researcher in selecting factor combinations when experimental units have four or more factors. Because STF predictions play such a vital role in experimental design, they will be discussed together. First we describe a questionnaire item that results from an experimental design.

Experimental designs produce factor level combinations. Each combination is a description of the system's capabilities. An example of such a description for the first experimental unit in Fig. 1 might read as follows:

> 30 percent of the important 2nd echelon force elements are identified in a timely fashion. Facilities can support 60 percent of the necessary immediate targeting activities. Tasking can be correctly communicated to 60 percent of the forces in time. There is timely access to the status of 10 percent of the Alert and Airborne forces. Weather data are three hours old.

Experts might be asked to judge how well they could perform their immediate targeting task in a command and control system that had these capabilities.

Next, we present examples of fully crossed factorial designs, which can be thought of as the "backbone" of other designs, and illustrate predictions that can be assessed from factorial designs.

Factorial Designs: Tests Between Interactive and Noninteractive Functions

In a fully crossed factorial design, every level of every factor is combined with every level of every other factor in the design. An example of a two-way factorial design of the Alert and Airborne Forces factors shown in Fig. 1 and described in Table 1 is shown in Panel A of Fig. 9. There are four factor levels for each of the two factors, so this 4 × 4 design produces 16 cells. Each cell represents a situation that would be described in a questionnaire item. For example, the upper left-hand cell would represent the situation where the command and control system had timely information about 90 percent of their airborne and alert forces.

A complete three-way factorial design of the alert forces, airborne forces, and weather factors shown in Fig. 1 and described in Table 1 is shown in Fig. 9B. Again each of these factors has four factor levels; this 4 × 4 × 4 design produces 64 cells or questionnaire items. For example, one item would describe a command and control facility that has timely access to 90 percent of their alert and airborne forces, and weather information that is only 15 minutes old. As the factors and factor levels increase, the questionnaire items generated from the fully crossed design increase rapidly.

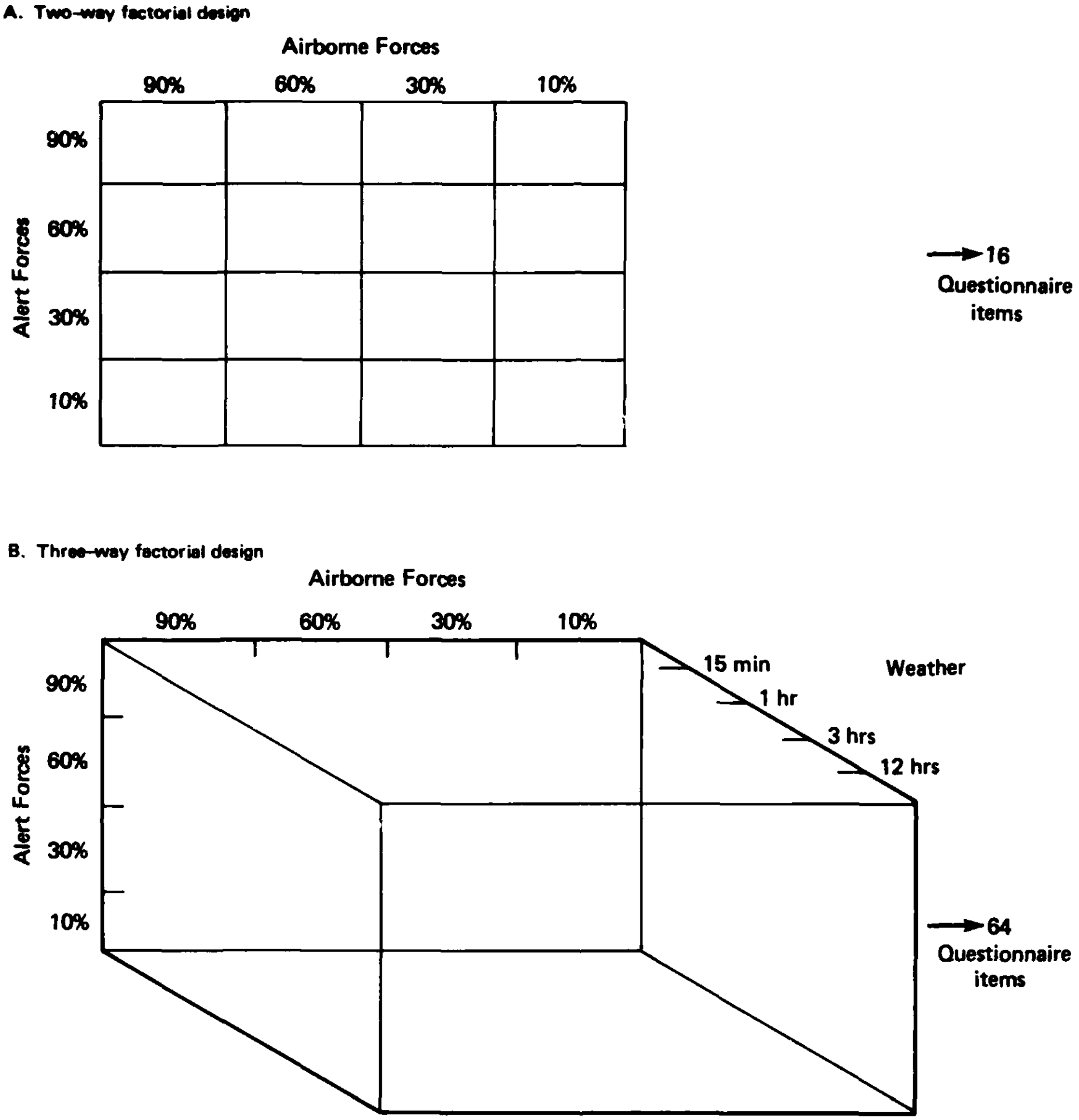


Fig. 9—Example of factorial designs

Factorial designs are useful for assessing main and interaction effects among factors.[1] When a proposed factor has no effect on judgments (after repeated tests), its appropriate parameter (either its weight or the scale value as the data and STF indicate) is set to zero in the STF. Tests of interaction effects provide a basis for choosing between noninteractive and interactive functions (panels A and B of Table 2) for those factors involved in the test. The hypothetical data shown in Fig. 10 illustrate this. For the two examples shown in Fig. 10,

[1] Manipulating the levels of a single factor in a one-way design will also provide tests of main effects. The factorial combination of factor levels provides an additional test of interactions among the factors included in the design.

assume that the J function in Eq. (3) is linear (this assumption is discussed in the section on scale-free tests).

The data from either panel A or panel B of Fig. 10 could have been obtained from the 4 x 4 design in Fig. 9A. Data are displayed graphically in Fig. 10 for diagnostic purposes. Mean response[2] is plotted as a function of one of the factor levels (Airborne Forces) with a separate curve for each level of the other factor (Alert Forces). These data exhibit a large interaction, an analysis of variance test of which would be significant. However, the F test would tell you nothing about the nature of the interaction, which is critical for deciding if the interaction should be interpreted, how it should be interpreted (what the experts are trying to communicate about the variables), and which function(s) out of many might account for the data. Figure 10A provides information about these issues. First, it reveals that the interaction is systematic; it is not the result of a lot of scatter (which might cause points on different curves to cross) or outliers.[3] Second, the interaction is divergent; respondents are saying that the less known about the status of the Airborne Forces, the *less* of a difference it makes how much is known about the status of the Alert Forces and vice versa. Both the range and multiplicative models shown in Table 2 can account for data that look like this. It is possible to choose between these two interactive functions through additional experimental designs and graphic diagnostics that will be illustrated later.

What if the data from the design shown in Fig. 9A turned out like those shown in Fig. 10B? The parallel form of these curves is what would be predicted by noninteractive functions such as those shown in Panel A of Table 2.[4] An analysis of variance test of these data would yield a nonsignificant interaction. Again however, the statistic does not provide a diagnostic for the researcher. In some cases, graphic inspection reveals small, systematic interactions. Systematic trends in the data should not be ignored, because the goal is to capture and measure such trends. In those cases, both interactive and noninteractive functions should be entertained to explain the data. However, the data in Fig. 10B are perfectly additive (noninteractive). Both additive and averaging functions can account for the observed parallelism. Additional experimental designs and graphic diagnostics (illustrated in the next section) are needed to distinguish between these two types of functions.

Tests among proposed interactive and noninteractive functions become more powerful as more factors and factor levels are included in the design as can be seen by a comparison of the lower two panels with the upper two panels in Fig. 10. The hypothetical data shown in Figs. 10C and 10D are for two levels of each factor. There are many more ways in the larger designs (Panels A and B) for curves to exhibit nonparallelism. Therefore, the idea that one of the class of additive functions is the appropriate combination process is more convincing when parallelism is observed. Similarly, when *systematic* interactions are observed with larger designs, it becomes more convincing that the observed systematic trends should be interpreted.

Factorial designs provide a basis for discerning interactive and noninteractive (additive) effects. However, they are not sufficient for distinguishing among models within either class. Next we discuss how graphic analyses combined with appropriate experimental designs can be used to rule out theories within the class of additive or interactive functions.

[2] Responses across experts would be averaged only for those respondents exhibiting the same divergent interaction. Data could also be plotted for individual respondents.

[3] Undue scatter and outliers tend to occur when respondents have not been sufficiently warmed up on the task or when questionnaires are formatted such that some item comparisons can be easily overlooked.

[4] The independence assumption associated with the class of additive functions predicts that curves should be parallel, not linear. That is, the vertical distances between the points on any two curves should be the same, independent of the value on the x-axis.

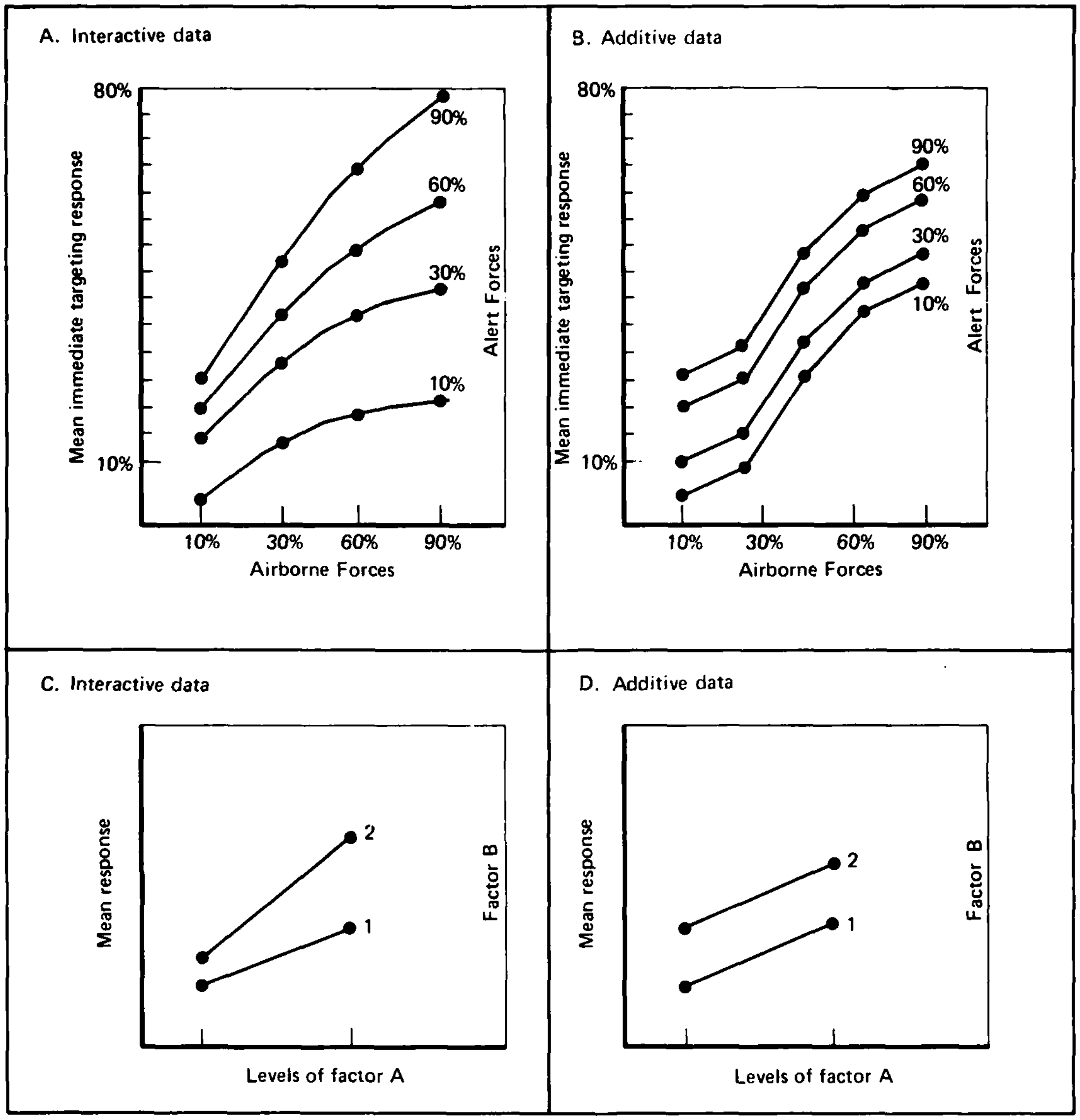


Fig. 10—Hypothetical data that might be obtained from 2-way factorial designs (panels A and B correspond to the 2-way design shown in Fig. 9, panel A; Panels C and D would be data obtained from 2 × 2 designs)

Factorial Design Extensions Permit Additional Tests of Hypotheses

One extension of factorial designs provides the researcher with greater diagnostic capability for distinguishing among functions; a second provides a test of the form of J shown in Eq. (3) relating subjective to objective responses.

Designs that Vary Amount of Information. Experimental designs must vary the amount of information contained in the questionnaire items in order to test among more of the unique predictions of functions shown in Table 2. Figure 11 presents a complete experimental design for three factors (Alert Forces, Airborne Forces, and Weather) that varies the amount of information. In this design, every possible item with one, two, or three pieces of information has been included; that is, every possible one-way and two-way factorial design has been included along with the three-way factorial design. Each cell in the first three one-way matrices produces an item with just one piece of information. An example of an item containing one piece of information from the first one-way design might read:

> The status of 60 percent of your Alert Forces is known.

An item from the third two-way matrix might read:

> The status of 30 percent of your Alert Forces is known.
> The reliable weather information is one hour old.

An item from the three way design might read:

> The status of 10 percent of your Alert Forces is known.
> The status of 90 percent of your Airborne Forces is known.
> Your reliable weather information is three hours old.

The complete design for three factors generates 124 questionnaire items. When items vary in amount of information, respondents are sometimes instructed to assume a baseline (current, planned, or otherwise defined) capability level for factors not presented. The decision to include this instruction depends on how reasonable the task seems to the respondents with the instruction omitted.

The graphs of hypothetical data presented in this section illustrate how the design feature of varying amount of information assists in diagnosing among functions. If data were obtained for the entire design shown in Fig. 11, it would be possible to test between a multiplicative and range function if the data were interactive as in Fig. 10A. It would also be possible to distinguish between additive and averaging models if the data were noninteractive as in Fig. 10B. The different predictions of these models are illustrated in Fig. 12. The hypothetical data plotted are the same as those shown in Figs. 10A and 10B, except for the dashed curve; data for the dashed curve would have been obtained from a one-way design of the factor plotted on the x-axis.

For each panel in Fig. 12, the relationship between the dashed curve (data that would be obtained from a one-way design) and the other curves in the figure (data from a two-way design) represents the *prediction* of the function written in the upper left-hand corner. In Fig. 12A, the multiplicative function predicts that the dashed curve should follow the same increase (or decrease) in slope that would be expected from the family of curves. This prediction can be seen from the algebraic formulation of the multiplicative function. When two factors are presented for judgment, this model predicts that the response should follow the form

$$r = A_i B_j . \qquad (4)$$

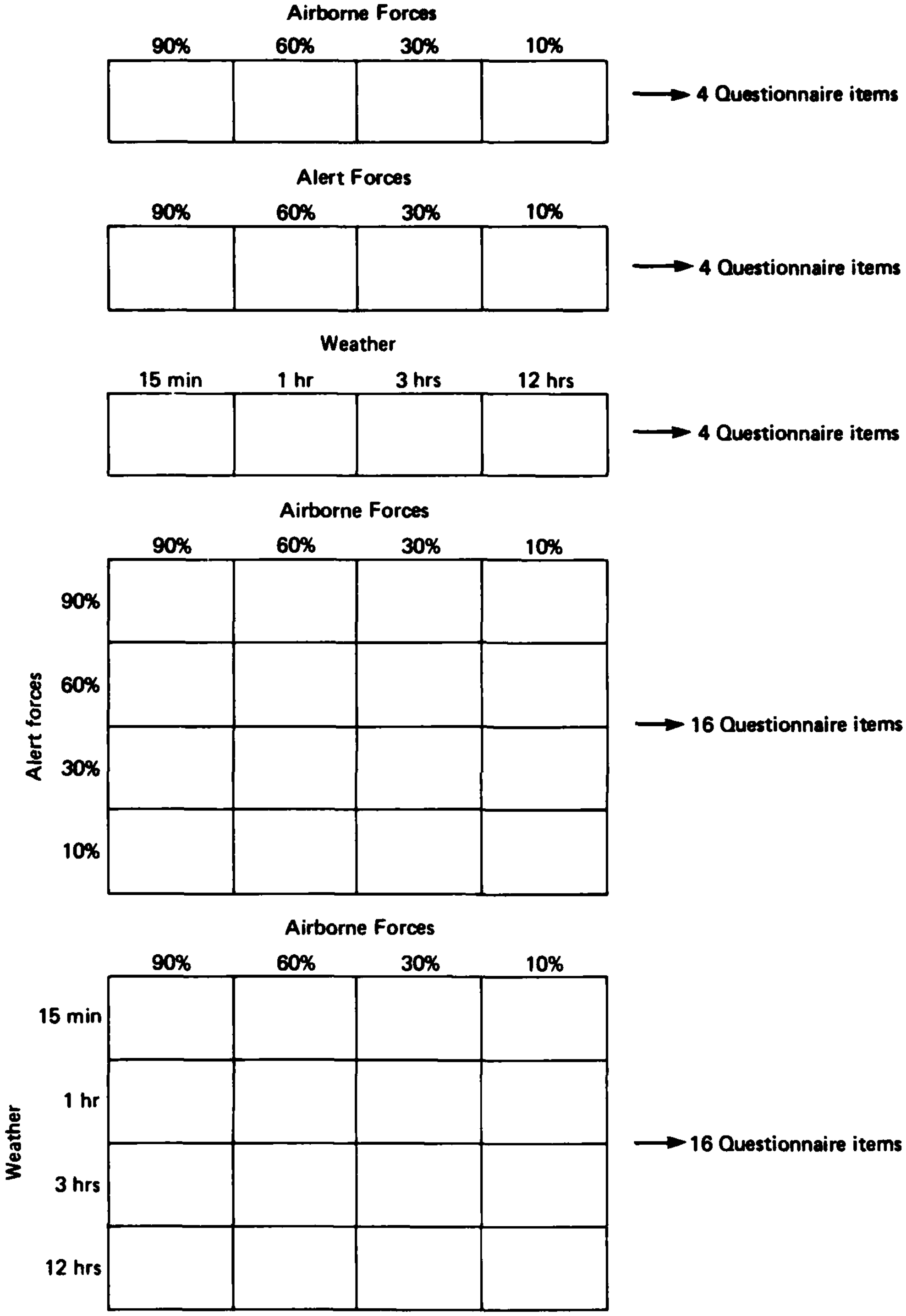


Fig. 11—Experimental design that varies amount of information (design is for three factors)

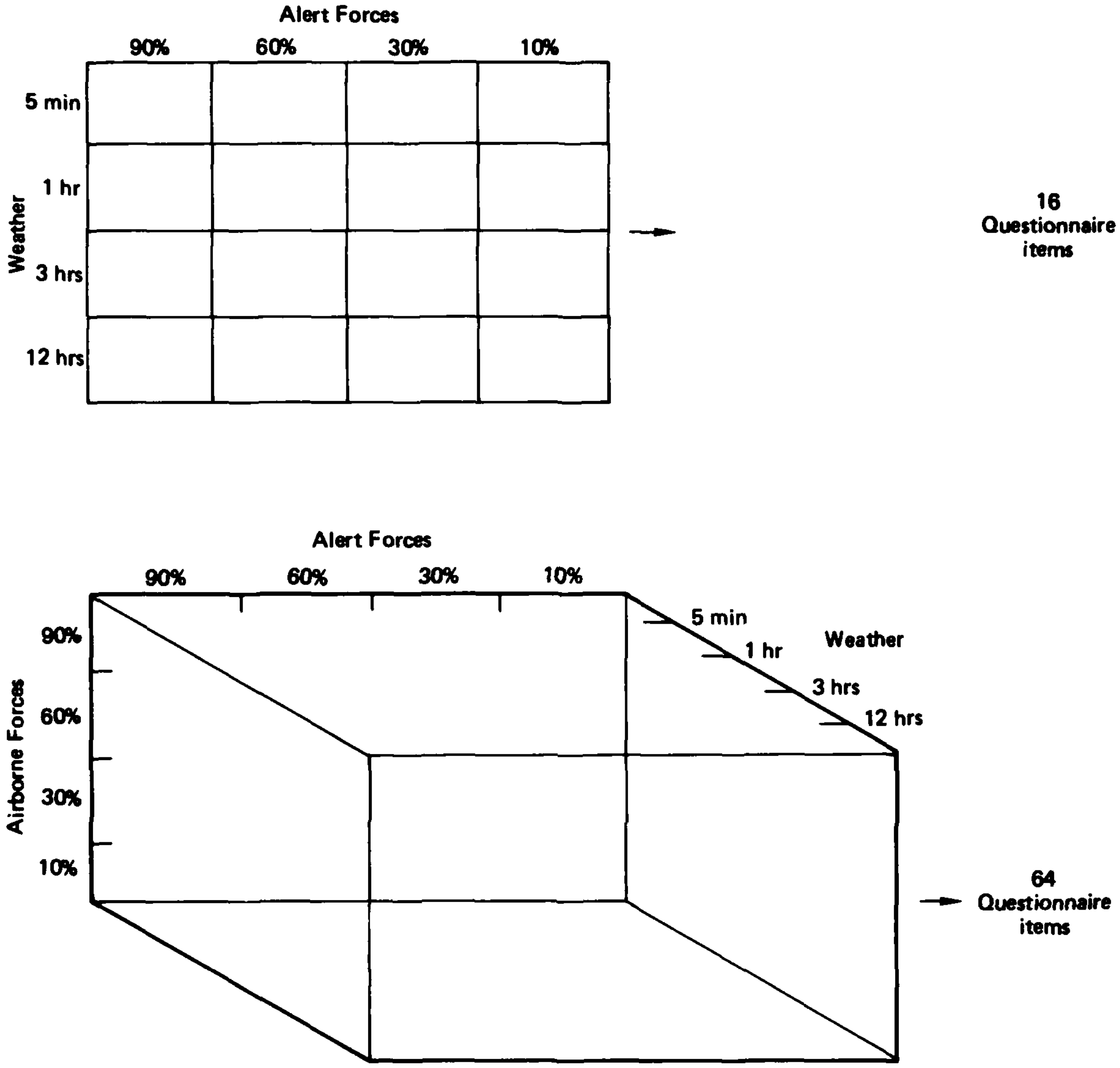


Fig. 11—continued

When factor A is presented alone, the function predicts

$$r = A_i x \, , \tag{5}$$

where x is the value of missing information. The two values are still multiplied and the curve that is obtained with only one piece of information should have a slope that follows the slopes of the family of AB curves. The *height* of the dashed curve (A alone) provides an indication of the value of the missing information (the value associated with the missing Alert Force information in this example).

If a range function is the appropriate function (Fig. 12B), the dashed curve should have a steeper slope than any of the other curves. Again, this can be seen from the algebraic form of

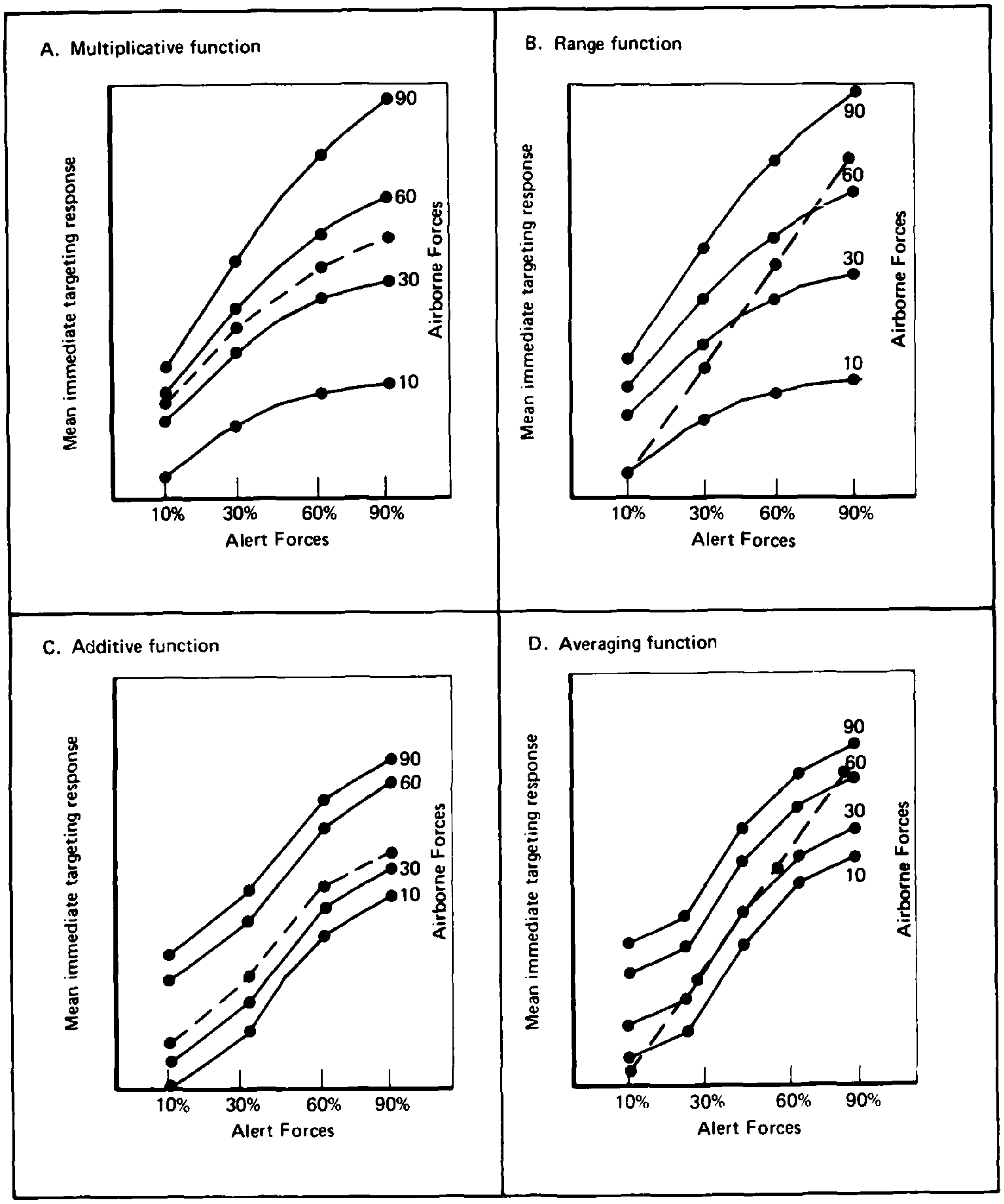


Fig. 12—Predictions of four different functions when designs vary the amount of information (dashed line represents data for the alert force factor presented alone)

the range function. When two factors (A and B) are presented for judgment, the function predicts that the response should follow the form

$$r = \frac{w_0 s_0 + w_A s_{A_i} + w_B s_{B_j}}{w_0 + W_A + w_B} + \omega(s_{max} - s_{min}) \ . \tag{6}$$

When only one factor (e.g., factor A) is presented for judgment, the weight of the missing information goes to zero and the response should follow the form

$$r = \frac{w_0 s_0 + w_A s_{A_i}}{w_0 + w_A} \ . \tag{7}$$

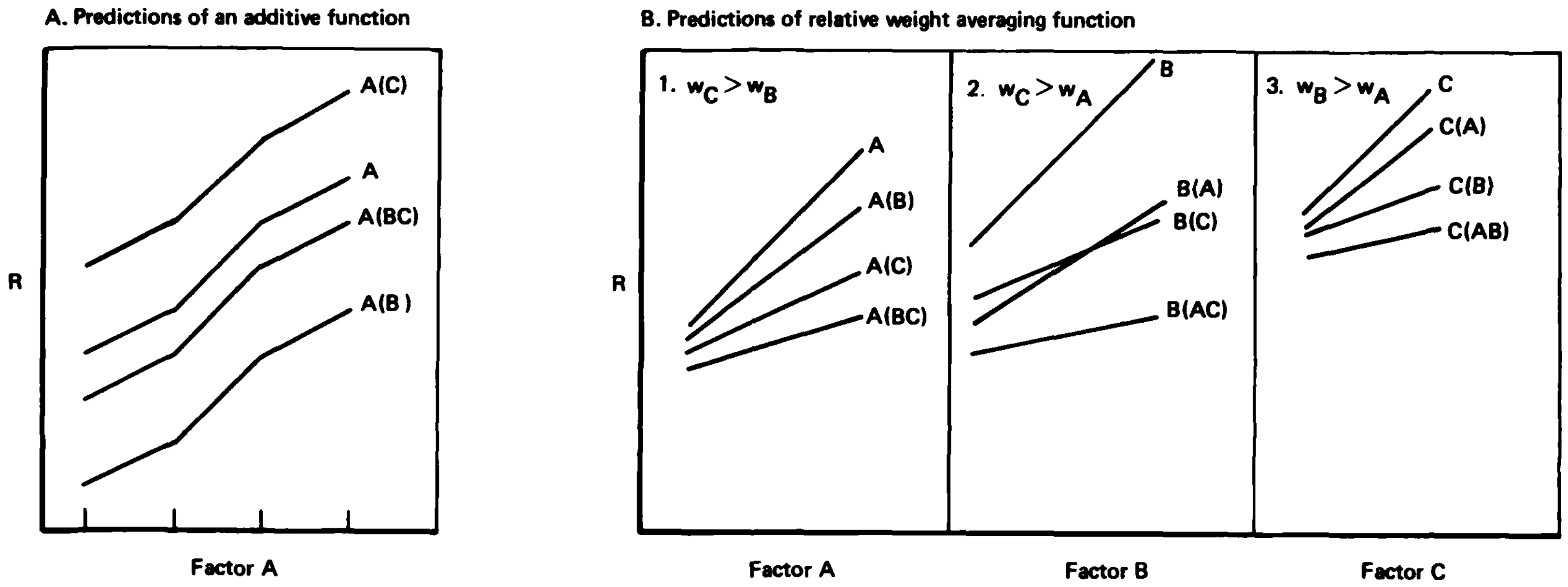
The denominator has decreased and thus the slope of A should increase, as shown by the dashed curve in Fig. 12B.[5] Again, the height of the dashed curve provides an indication of the value of the missing information. The same reasoning follows for distinguishing between an additive and an averaging function. The different predictions of these two functions are shown in Figs. 12C and 12D.

The design illustrated in Fig. 11 provides stringent tests among these four functions. The tests are stringent because the design offers *repeated* opportunities for the predictions of a proposed function to fail. For example, if an interactive interpretation is appropriate for two factors (A and B), an interaction between them should be observed in both the two-way and the three-way design. If the multiplicative function is the appropriate interactive function, a graph with A on the x-axis with a separate curve for each level of B together with A alone should look like the graph with B on the x-axis with a separate curve for each level of A together with B alone; they should both look like Fig. 12A. If interactions were observed among all three factors (A, B, C) and a multiplicative function was appropriate, the form shown in Fig. 12A should be observed for all factor pairs as well as for graphs of all three factors (e.g., the BC design plotted on the x-axis with a separate curve for each level of A together with BC alone). These same plots should follow the forms shown in 12B, 12C, or 12D for a range, additive, or averaging function, respectively.

Other graphic diagnoses of data are possible when the design shown in Fig. 11 is used. For example, the additive function shown in Table 2A predicts that the effect of a factor should be independent of the factors with which it is paired. This prediction can be assessed graphically for each factor separately. Figure 13A illustrates this for factor A. The top curve in the figure represents data that would be obtained from the AC design averaged over C; the next curve down, the data from factor A presented alone; the next curve down, the data that would be obtained from the ABC design averaged over B and C; and the bottom curve from the AB design averaged over B. All the curves have the same slope, which is predicted by the additive model shown in Table 1A.

An averaging model (e.g., the relative-weight model shown in Table 2A) predicts that the effect of a factor depends on the number of other factors it is paired with. Thus, the slopes of the same curves should vary with amount of information contained in the item. This prediction can be seen from the curves in Fig. 13B. Each set of four curves is for a different factor. The curves in Panel B1 would be data from the four designs in which factor A is included (the A alone—top curve—AB, AC, and ABC designs). Similarly, the curves in Panels B2 and B3

[5] In the range function, the effect of the missing information is incorporated in the relative weight of the initial impression, s_0, which is greater when information is missing. The range term drops out in Eq. (7), because only one factor is presented for judgment; the relative weight of S_0 is $w_0 \ \ (w_0 + w_A)$.



NOTE: A: Data from the one way design for factor A;
A(C): Factor A averaged over all levels of factor C in the AC design;
A(B): Factor A averaged over all levels of factor B in the AB design;
And so forth for the other curves

Fig. 13—Graphic predictions of additive and averaging functions (examples are for three factors, A, B, and C, combined according to the experimental design shown in Fig. 11)

would be data from the B alone, AB, BC, ABC, C alone, CA, CB, and CAB designs. The slopes of each set of curves not only reveal interactions (which would violate an additive function) but they reveal the *order* of the weights of the averaging function. The curves in Panel B1 say that the weight of factor C is greater than the weight of factor B, because its slope is *less*. This prediction can be seen from the functional form of the relative weight function. The slope of A when paired with just C is

$$\frac{w_A}{w_0 + w_A + w_C}\,, \tag{8}$$

because the judgment would follow the model

$$r = \frac{w_A s_{A_i}}{w_0 + w_A + w_C} + \frac{w_0 s_0 + w_C s_{C_k}}{w_0 + w_A + w_C}\,. \tag{9}$$

However, the slope of A when paired with B would be

$$\frac{w_A}{w_0 + w_A + w_B}\,. \tag{10}$$

The only difference in the two slopes in Eqs. (8) and (10) is the weight of factor C in Eq. (8) versus the weight of factor B in Eq. (10).[6] Using the same reasoning, one would conclude that the weight of factor C is greater than the weight of factor A from the order of the curves in Panel B2 of Fig. 13, and that the weight of factor B is greater than the weight of factor A from the order of the curves in Panel B3. Thus the experimental design shown in Fig. 11 would permit a test of transitivity of the values of the weighting parameters in the averaging models.[7]

The purpose of experimental designs is to test unique predictions of algebraic functions under investigation. When the predictions of a particūlar function are repeatedly observed in the data, that function becomes more and more credible as the appropriate function to explain the data. When the data do not follow the predictions of any of the hypothesized functions, they would all be rejected as appropriate theories of the data.

Scale-Free Designs: Tests of J. The scale-free design (Birnbaum, 1974; Birnbaum and Veit, 1974b) was developed to resolve the measurement problem of when it is appropriate to transform observed interactive curves such as those shown in Fig. 10A to parallel curves (e.g., Fig. 10B). There are two conflicting schools of thought in psychological measurement about how to handle observed interactions. One is to assume that the function relating observed to subjective responses (J in Eq. (3)) is only monotonic. This view is based on the idea that only ordinal (not metric) information is contained in responses and that an additive function is most appealing because of its "simplicity." Therefore, if interactions are observed in the data and a monotone transformation[8] can be found to transform the interactions away (transform the nonparallel to parallel curves), this would be the "appropriate" procedure and tests would be among additive functions. (Of course if curves were not monotonically related

[6] If the weights of factors B and C were equal, the slopes of the A(B) and A(C) curves would be the same. This is why the slopes of curves from designs that *vary* information (e.g., A, A(B), and A(BC) in Panel B1) need to be compared in order to differentiate between an additive and an averaging function.

[7] To derive weight independently from scale value parameters in models where these two parameters form a product, it is necessary to vary the amount of information in the experimental design. This problem is discussed and illustrated further in Norman (1976) and Birnbaum and Stegner (1981).

[8] A monotone transformation changes the relationships among the data points while maintaining their original order.

to a set of parallel curves, the observed interactions would be interpreted—tests would be among interactive models.) Another approach is to assume that metric information is contained in the responses (the J transformation relating r to R in Eq. (3) is linear) and thus to explain observed interactions. The scale-free design is a method for *testing* which of these assumptions is appropriate. That is, the design provides a basis for deciding if observed interactions reflect the underlying *subjective* process (T in Fig. 8.).

The importance of such a design feature can be viewed with respect to Table 2. If interactions are observed in the data and are interpreted as perceptual, tests would be among interactive functions (e.g., those described in Table 2B); subjective scale values would be obtained from the interactive function that best explained the data. However, if the interaction was transformed away, tests would be among noninteractive functions (those described in Table 2A); a *different* function *and* set of subjective scale values would be required to explain the transformed data.[9] *Both* interpretations of the data cannot be "right." The scale-free test provides a testable basis for deciding whether to transform away observed interactions.

The basic idea of the scale-free test is that, by embedding two combination processes in one task, a test of the interaction among the factors for one of the combination processes is possible without making any assumptions on the form of the function relating observed to subjective responses (J in Eq. (3)) for that process. Take, for example, the question, "Do the factors Emitter Location and Emitter Coverage (Fig. 1) combine additively in the combination process?" An example of a task that requires two combination processes would be to have targeteers *compare* two command and control systems in terms of the percentage of important enemy targets they could identify; each system would be described by an emitter location *and* a coverage capability. The first combination process requires that respondents combine the enemy emitter and location capability information that describes each system; the second process requires that they *compare* the two command and control systems (e.g., judge how much better they could perform their targeting task with one system than with the other). The hypothesis being tested might be that the first process follows an additive combination function and the second process a comparison (subtractive) function as instructed. This hypothesis can be written

$$R_{ijkl} = J[(L_i + C_j) - (L_k + C_l)] , \qquad (11)$$

where R_{ijkl} is the observed comparison response, L_i, C_j are the ith and jth levels of Emitter Location and Coverage making up one command and and control description, L_k and C_l are the kth and lth levels of Emitter Location and Coverage, respectively, making up the other command and control description, and J is some monotonic function relating subjective to observed comparison responses.

The design that provides a scale-free test of additivity for Location and Coverage (and hence a test of Eq. (11)) requires a factorial design of the Location and Coverage factors *and* of the Location/Coverage combinations being compared. This scale-free design is illustrated in Fig. 14. For three factor levels of Emitter Location and four factor levels of Emitter Coverage (Table 1), a complete 3 × 4 (Location × Coverage) factorial design would produce 12 Location/Coverage combinations as shown in Fig. 14A. These 12 combinations are then used to form a 12 (rows) × 12 (columns) factorial matrix. Each cell of this design contains descriptions of *two* command and control systems; each system is described by a Location and Cover-

[9] A full explanation of the data requires specification of both T and J in Fig. 8 and Eqs. (2) and (3). It is questionable whether the explanation that T is additive and J is some nonlinear monotonic function is "simpler" than the explanation that T is nonadditive and J is linear.

A. Factorial design of location and coverage factor levels

Location	Coverage 90%	60%	30%	10%
10 m	L/C$_1$	L/C$_2$	L/C$_3$	L/C$_4$
100 m	L/C$_5$	L/C$_6$	L/C$_7$	L/C$_8$
1000 m	L/C$_9$	L/C$_{10}$	L/C$_{11}$	L/C$_{12}$

→ 12 Location/Coverage combinations

B. Factorial design of location/coverage combinations

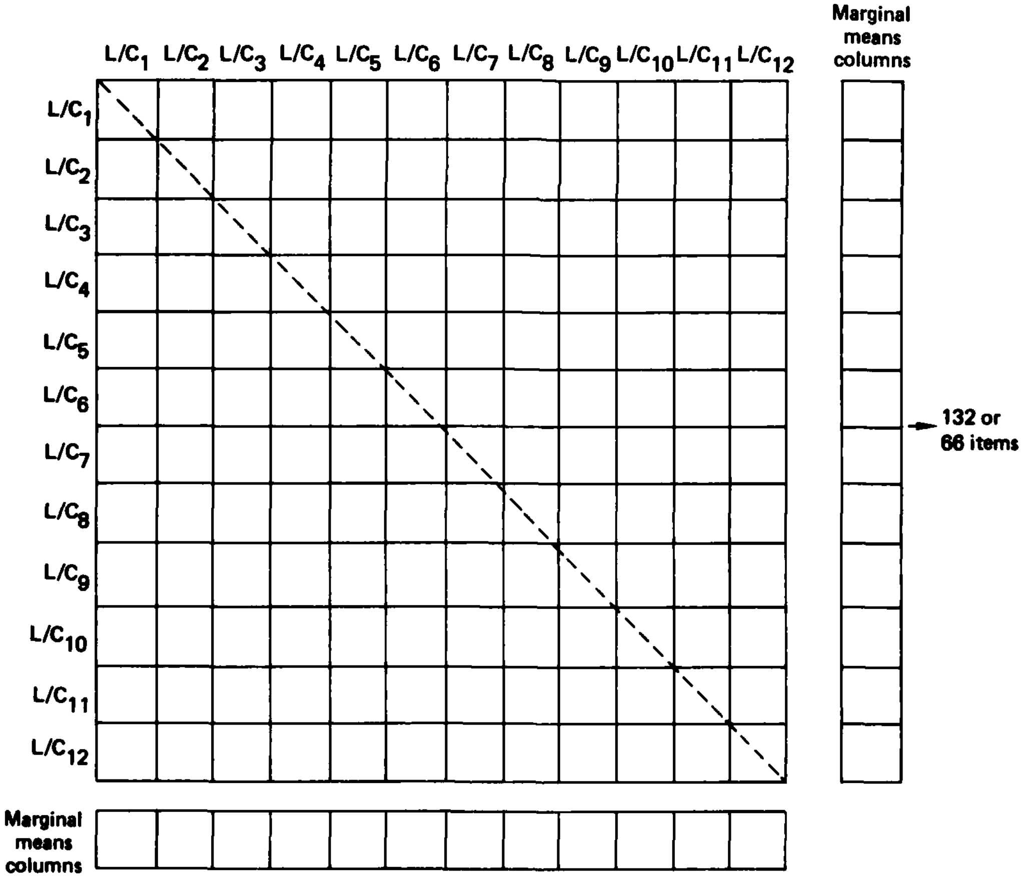


Fig. 14—Scale-free design for the enemy emitter location and coverage factors

age capability. This symmetric matrix shown in Fig. 14B contains 144 cells. The diagonal cells contain identical combinations and therefore would probably be eliminated from the questionnaire,[10] leaving 132 comparisons. Because the upper and lower triangles of this matrix contain the same combinations, one triangle could be omitted from the questionnaire, if time doesn't permit repetitions, leaving 66 comparisons. However, when triangular designs are used, it is assumed that the scale values for the row and column stimuli are equal. This may not be an appealing assumption in situations where there is reason to believe that factors not manipulated in the experimental design, such as order of item presentation, may differentially affect row and column stimuli. Reasons for avoiding triangular designs are described in Birnbaum (1981).

The marginal means of the large 12 × 12 data matrix shown in Fig. 14B are the *subjective* scale values of the Location/Coverage combinations *under* a subtractive model.[11] Thus, they are the scale-free values needed to evaluate the Location × Coverage interaction. These values are referred to as "scale-free" because they do not depend on the form of the J function relating observed to subjective *comparison* responses. They depend only on the fit of the subtractive function to the data. J in Eq. (11) is assumed to be only monotonic, so any perturbations found in the comparison data can be scaled away to provide better estimates of the scale-free values.

A graphic test of the Location × Coverage interaction can be obtained by placing the scale values (marginal means)[12] in their appropriate cells in the 3 × 4 matrix (Fig. 14A) and plotting them as shown in Fig. 10. If systematic interactions are observed in the resulting curves, interactive functions would be entertained as the explanation of the effects of these two factors in judgments of targeting ability. The results would argue for interpreting (not scaling away) future observed interactions between these two variables. Parallel curves would dictate entertaining additive functions and scaling away interactions when they are observed between these two variables.

The major advantage of using a comparison task in the scale-free design is that a subtractive function has a good track record in accounting for comparison judgments on a variety of judgment dimensions (Birnbaum, 1974, 1980; Birnbaum and Veit, 1974a,b; Rose, 1980; Rose and Birnbaum, 1975; Veit, 1978; Veit, Rose, and Ware, 1982). However, another task or function could be entertained. As long as the data were monotonically related to the predictions of the functions, the data could be transformed in accord with the predictions and the scale-free values obtained from the function. In the scale-free approach, observed responses need be only monotonically related to psychological responses to get a scale-free test of the *embedded combination* process.

The scale-free design calls for embedding factorial designs. Questionnaire length increases substantially with the addition of more factors. Below we discuss guidelines for reducing the number of items fielded for study while maintaining sufficient constraints for testing proposed STFs.

[10] Numbers defined in the response scale as "equal" (e.g., in being able to identify important targets) might be put into these cells for data analysis.

[11] In a complete factorial matrix, the row marginal means are linearly related to the subjective scale values of the row factor levels and the column marginal means are linearly related with a negative coefficient to the column scale values under a subtractive function (e.g., Eq. (4)) when instructions are to compare row with column factor level (row minus column).

[12] The row and column marginal mean for a given Location/Coverage combination would be averaged when they are close to the same value.

SELECTING EXPERIMENTAL DESIGNS

When factors are being explored for the first time and a decision must be made as to the appropriate STF (and hence subjective values), it is important to obtain information on interactions among factors using the scale-free design as well as to vary amount of information contained in items (Fig. 11), so that it is possible to test among alternative functions. However, if very many factors are being investigated (e.g, the six factors shown in experimental unit 1 of Fig. 1) and each has between three and four factor levels, it becomes impossible from a practical standpoint to field a questionnaire that includes a complete set of items from these designs.

Two techniques in combination make it possible to to field reasonably sized (up to 200 items) questionnaires: (a) gathering judgment data in stages and (b) selecting a subset of a complete array of experimental designs. The first technique requires that pilot studies be conducted prior to a final STF testing stage. These would provide preliminary information about the effects of the factors on judgments. The second technique focuses on an experimental design selection process, the goal being to reduce the questionnaire length while maintaining the constraints necessary to test adequately among proposed STFs.

Pilot Study Phase

A preliminary investigation of the factors under consideration aids greatly in (a) assessing if the judgment tasks are feasible to the respondents, (b) reducing the number of structure/function hypotheses to be tested in the STF testing phase, and (c) providing a more solid base for conclusions concerning appropriate STFs through repeatability of results. In the pilot study phase, emphasis might be on main effects of factors, distances between factor levels, scale-free interactions among the factors, and testing predictions of the STFs under investigation. In this phase, different questionnaires, each addressing a different question about the factors under consideration, could be fielded to two or three respondents within an expert group. The amount of information obtained in this phase will depend on time, resources, and availability of respondents. Often, adequate determination of experimental questions requires fielding more than one questionnaire for an experimental unit, perhaps because questions generated by results from a first fielding need to be answered, results are not clear, or more information is desired about the relationships among the factors.

The following experimental design descriptions offer some guidelines for a first round of questionnaires. All factors and factor levels should be included in some aspect of the experimental design, possibly by including all two-way factorial designs or a mixture of two- and three-way designs, depending on the size of the questionnaire that is generated. Four-and five-way designs might be included to get an idea of higher-order interactions. For these larger designs, the number of factor levels could be reduced to two. (When reducing the number of factor levels for a given design, it is a good idea to include in the selection what is believed to be the highest and lowest valued level of each factor so as to span the full range of the factor dimension.) Scale-free designs could be included for some factor pairs. Exactly how much is fielded in a first round depends on how many respondents would be available.

Results from a first round of questionnaires may suggest several changes in the factors that require more data collection. First, preliminary results may indicate that a factor should be redefined. Indications may result from verbal reports of confusion about the factor's definition, no effect of the factor on judgments, or individual differences in the effects of the factor on judgments (for example, the ordering of the factor levels or the direction of a factor's interaction with other factors might be different for different respondents).[13] Second, preliminary results may reveal very small differences among some of the factor levels. If *new* levels were hypothesized, these should be tested. Third, factors that displayed interactions would be refielded in scale-free designs, if this had not been done in the first round. Some of these could be embedded in three-way designs, again guided by the size of the questionnaire generated by the design and the number of respondents available.

Results from the pilot study guide the shape of the experimental designs used in the STF testing phase in the following way: If a factor had no effect on judgments, it would be dropped from further study.[14] Factor levels close in value would be replaced with only one level. Scale-free tests would provide information about whether factors combined interactively or noninteractively and thus reduce the number of STFs that need to be taken into account in the final design.

STF Testing Phase

In this phase, it is necessary to use experimental designs that allow adequate tests of the STFs under consideration in each experimental unit. This design will lead to the final conclusions about the appropriate STFs. As illustrated earlier, a major experimental design feature would vary the amount of information contained in an item (see Fig. 11). When more than three factors are included in a unit (assuming between three and five levels of each factor), a complete design of all possible numbers of factors (all singles, pairs, triplets, etc.) would yield an impractically long questionnaire. Thus, it is necessary to select a subset of factorial designs from a complete array of designs (all possible one-way, two-way, three-way, four-way, etc., depending on the number of factors defining the experimental unit) that allow sufficiently *stringent tests* of the STFs under consideration. Selection of factorial design subsets should be based primarily on what is known about the unique predictions of the most viable STFs, which themselves would be based on information obtained in the pilot study. The following provide some guidelines for selecting or excluding design subsets.

1. Reduce the number of factor *levels* in the larger designs (e.g., three-way or larger) for those factor combinations that received more stringent tests (all levels were used) in the pilot study. Levels at the top and bottom of the factor level ranges should be retained.

2. When four or more factors are included in the experimental unit, use the strategy of confounding factors (not fully crossing each factor with every other factor) to generate items that include a factor level of each factor; make sure factors that are confounded in larger designs are unconfounded (fully crossed) in smaller (two- or three-way) designs. Decisions about what factors to confound would be based on pilot data. For example, factor pairs investigated more thoroughly in the pilot phase might be selected to confound in the larger designs in the STF testing phase. Also, factors that are correlated in the real world might be selected for confounding in the experimental design. Examples of confounded designs are presented for five factors in Birnbaum and Stegner (1978), and for the six and seven factors of experimental units 1 and 2, respectively, shown in Fig. 1 (Veit, Callero, and Rose, 1982).

[13] A factor definition that produced individual differences among respondents would not be considered appealing because it would require the complicated conclusion of more than one STF at that path of the structure, or more than one set of parameters for a given STF at that path. If the choice were between a factor definition that produced agreement and one that did not, the one that produced agreement would be selected.

[14] If the capability defined by the factor was important because, for example, decisions were pending about developing the equipment that increased that capability, it could be retained in the structure but given either a zero weight or zero scale values in the STF.

As an alternative to confounding factors, each factor can be reduced to two levels (e.g., the top and bottom of the range) and combined in a 2^n design where n represents the number of factors defining the experimental unit. If feasible in terms of questionnaire length, this design provides a lot more information than confounding factors about the relationships among the factors.

3. Include enough subsets of one-way, two-way, etc. factorial designs to provide adequate tests of proposed STFs (e.g., those shown in Figs. 12 and 13). The adequacy of the tests are guided by what is known about the unique predictions of the STFs under consideration. The researcher should include factorial design subsets that provide a mathematically unique solution of the parameters of the STFs under investigation. That is, the subset of designs must produce a set of linear equations and unknowns for each STF under consideration from which to obtain a unique solution for the unknowns (one of the parameters can be fixed to some value without loss of information).[15] Other design subsets are necessary to provide a test among predictions of the STFs being hypothesized. Graphs of predicted results (such as those shown in Figs. 12 and 13) could serve as the selection guide. Of course, additional designs that make a viable function "work harder" are desired, because the more data an STF can account for, the more credible it becomes. Again, additions should be guided by what is known about the unique predictions of the STFs under consideration and the questionnaire length.

COLLECTING JUDGMENT DATA

Data collection can be broken down into selecting respondents and administering the questionnaire. The availability and professional characteristics of the respondent population need to be considered in formulating the structural hypotheses. The respondent population associated with each experimental unit should have credibility to those requesting the system evaluation.

Questionnaire administration consists of familiarizing the respondents with the task and having respondents fill out the questionnaire. Task familiarization or review is important for "setting the stage." The respondents should feel at home with the situations they will consider and completely understand the judgment task required of them. Before the questionnaires are administered, respondents are briefed on and discuss any necessary background information (e.g., the details of the battle for a command and control problem domain), the factor definitions, factor levels, and judgment task. The length of this preparatory session will depend on previous participation of the respondent group in structure development or pilot studies. After this session is completed, respondents complete from 10 to 20 items consisting of representative items to familiarize them with the task. Then they fill out the questionnaire.

[15] This may not be possible for some functions—for example, the multiple regression additive function.

IV. DETERMINING STFs AND FINAL STRUCTURE

After each data collection session, analyses are performed that provide tests among competing STF/structural hypotheses. The structural hypotheses revolve around the appropriateness of hypothesized intermediary factors (e.g., Fig. 3).

DATA ANALYSES

Determining Individual Differences

The first step in data analysis is to look at the data for each respondent separately to get an idea of their similarities and differences in the effects of the factors on judgments. Data within an experimental unit are combined when there are no differences in the ordering of their factor levels and the pattern of interactions (e.g., divergence, convergence) observed. When either the factor level ordering or the pattern of their data differs, the two sets of respondents would not be combined for analyses. One goal in the pilot study phase is to find factor definitions and task descriptions that affect the experts in similar ways, because it is simpler to have *one* STF at each path in the structure. If it is not possible to resolve these differences, research should be conducted to determine the bases (e.g., military rank, training background) of the observed differences and incorporate those differences as part of STF theory.[1]

Testing Among STFs

The explanatory power of a proposed STF lies in its ability to reproduce the systematic details of the data. In each experimental unit, we follow two major steps for testing among the abilities of the proposed STFs to do this. First, we rely primarily on graphic analyses to reduce the number of structure/function hypotheses to a select few. Second, we use additional graphic tests combined with the least-squares data-function discrepancy criterion to test among the remaining hypotheses. These steps are illustrated below. For purposes of discussion, assume that the scale-free tests fielded in the pilot stage have provided the basis for assuming J in Eq. (3) linear.

Graphic Tests Reduce the Number of Possible Hypotheses. Figures 12 and 13 have demonstrated how graphic tests provide a powerful diagnostic tool for distinguishing among algebraic functions. This drastically reduces the number of viable hypotheses. However, a few hypotheses are often retained as possible explanations of the data, especially in large designs (more than three factors) where many factor combinations that would clinch a particular hypothesis might not have been fielded because of attempts to reduce questionnaire length. (The more extensive the pilot study, the fewer ambiguities.) Two examples are presented below.

For the first example, say our experimental design had six factors (e.g., experimental unit 1 in Fig. 1). Label the factors A, B, C, D, E, and F. If the pattern of the data for factors A, B,

[1] Functional differences could be in the algebraic form of the STF, magnitudes or order of the factor weights, magnitudes of the scale values, any combination of these three areas, or all three areas.

C, and D followed that shown in Fig. 15B for the designs fielded, but the convergence was much smaller, and factors E and F followed the pattern in Fig. 14C, a viable hypothesis would be a range function with a positive ω weight of the range term for the first four factors and a multiplicative function for factors E and F. However, a small convergent interaction for factors A, B, C, and D might also lead to the retention of the relative-weight averaging function for these factors. The next question is, How might these two *groups* of factors combine? If a graph with EF on the x-axis with a separate curve for each ABCD combination looked like Fig. 15C, a multiplicative function would be supported.[2] The structure shown in Fig. 16A depicts

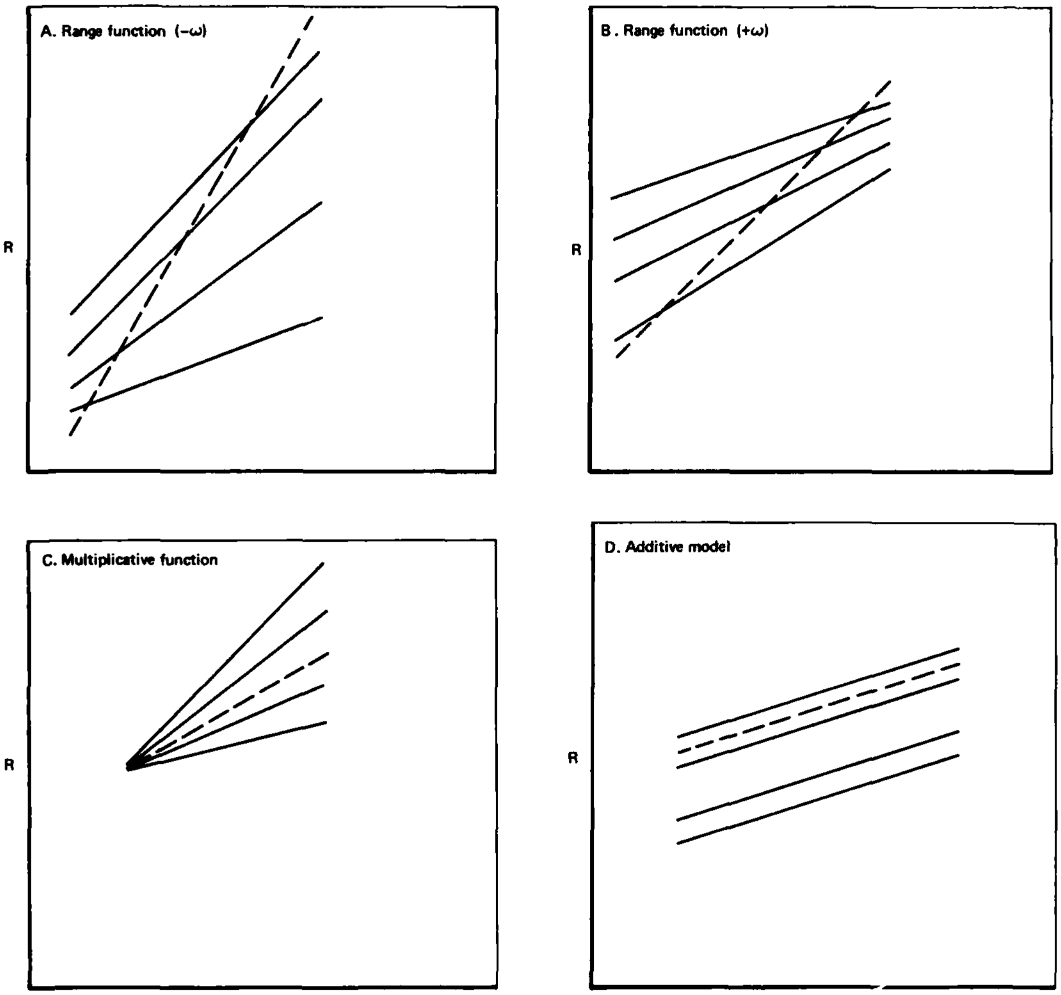


Fig. 15—Hypothetical data predicted by different algebraic functions (dashed curves represent the data for the factors plotted on the x-axis when presented alone)

[2] If the ABCD × EF factorial design was not fielded (or another design that contained this information—e.g., the ABCDEF design), it would be necessary to assess the relationships between the factors in these two groups using the designs at hand (e.g., the AE, BCF, DEF designs), which usually leads to the retention of more hypotheses.

the combination functions for the two factor sets, as well as the multiplicative combination function for combing the two sets; two STFs have been retained for the ABCD factor set.

Another example has five factors (A, B, C, D, and E) in the experimental unit. Graphic diagnostics revealed a divergent interaction (15A) for factors C, D, and E. However, again suppose the interaction appeared small. For such data, one might want to retain both a relative-weight averaging function and a range function with a negative ω. Suppose that graphic diagnoses also revealed a multiplicative relationship (Fig. 15C) between A and B and additivity (Fig. 15D) for the two factor sets (a graph of the AB × CDE data—AB on the x-axis with a separate curve for each CDE combination—looked like Fig. 15D). This might lead to the factor grouping shown in Fig. 16B, with two possible functions for the CDE factor set. However, all interactions in these data were divergent, so it would be desirable also to investigate a five-factor range model with a negative ω. This alternative is shown in Fig. 16C.

These examples demonstrate how preliminary diagnostics can reduce the number of structure/function combinations that should be retained as explanations of the data. The more that is learned about the relationships among the factors in the pilot study, the more alternatives can be ruled out.

Testing Among Selected Structure/Function Hypotheses. Tests among remaining structure/function combinations for each experimental unit are made using STEPIT, a parameter-estimation program (Chandler, 1969) that selects parameters that minimize the sum of squares discrepancies between the data and the STF's predictions. One can write each hypothesis under consideration into the program by embedding STFs associated with intermediary factors into the function at the outcome (suboutcome) path for the experimental unit. For example, one structure/function test for the five factors shown in Fig. 16B would be to embed a multiplicative function for the AB factors and a range function for the CDE factors in an overall additive function.[3] Another structure/function test would be a five-factor range function (Fig. 16C). The program provides a statistic of the sum of squares data/function discrepancy along with the parameter estimates of the function. The STF with the smallest discrepancy would be considered the "best-fit" STF for that experimental unit.

As mentioned earlier, if deviations are large and systematic for the statistically "best" function, that function would also be rejected as the appropriate STF. Graphs that plot both predicted values (r in Eq. (2)) and obtained values (R in Eq. (3)) on the y-axis for the different factorial designs used in the experimental unit provide a means for assessing the magnitude, direction, and systematic nature of data/function deviations. Such graphs aid in decisions to reject functions and in determining a "correct" function. If a new function suggested by the pattern of deviations cannot be adequately tested on the available data (the designs used are not sufficient for testing the newly proposed functions' unique predictions), it would be necessary to redesign the experiment and collect new data.

[3] This embedded function can be formulated as follows:

$$r_{AB} = s_{A_i} s_{B_j}$$

$$r_{CED} = \frac{w_0 s_0 + w_C s_{C_k} + w_D s_{D_l} + w_E s_{E_m}}{w_0 + w_C + w_D + w_E} + \omega(s_{max} - s_{min})$$

$$r_{ABCDE} = r_{AB} + r_{CDE}$$

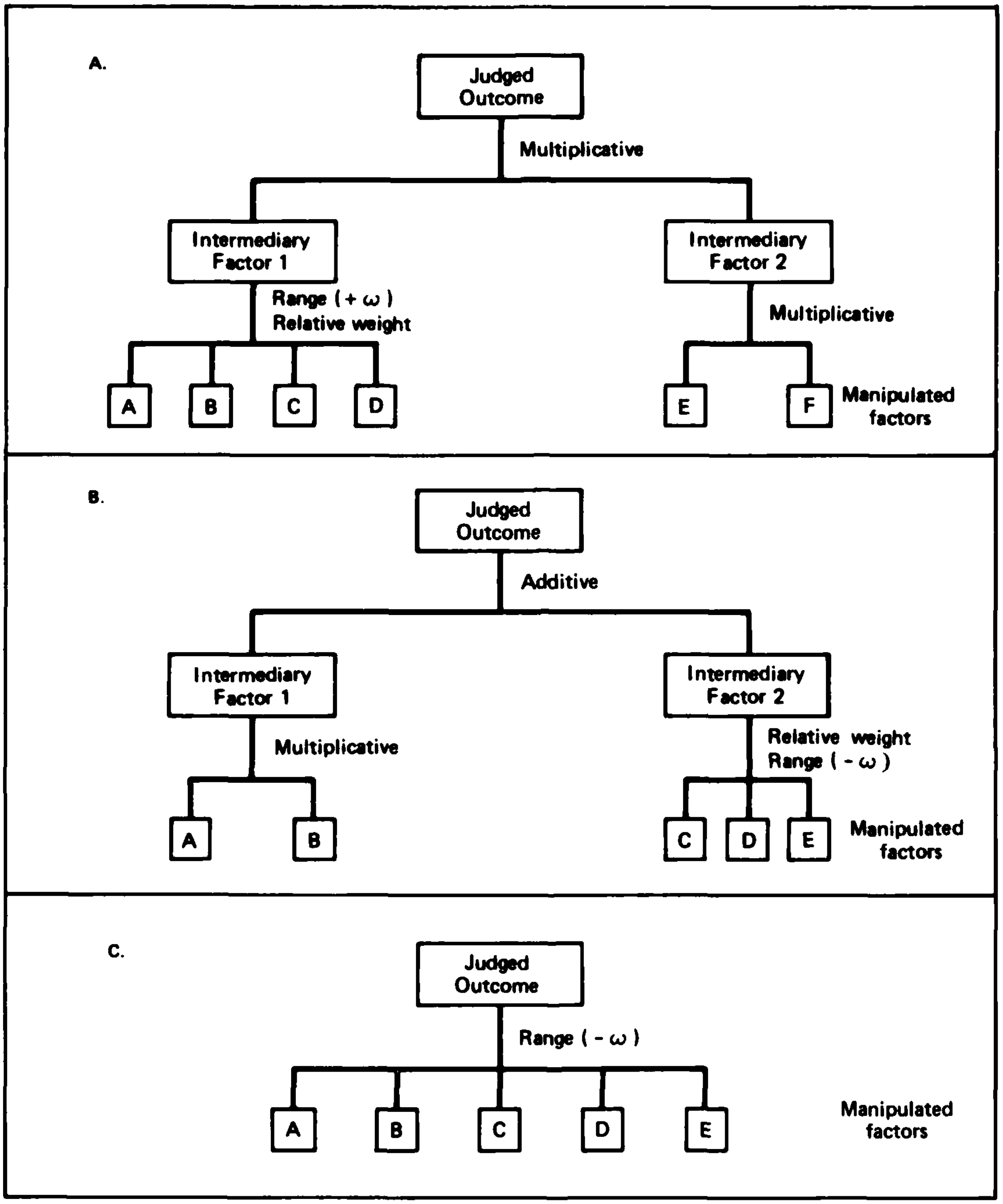


Fig. 16—Alternative structure/function combinations

Figure 17 is an example of a graph that has plotted both predicted and obtained values on the ordinate. The factors in Fig. 17 came from the top experimental unit shown in Fig. 7. Respondents judged the "chance of winning the Land Battle," given the number of enemy Fixed Targets, Moving Force Elements, and Stationary Force Elements neutralized by friendly forces. Circles in each panel represent the mean judgments of "chance of winning the Land Battle" and the curves represent the predictions from the range function shown in Table 2B. The data/function deviations are very small. Most of the circles are falling right on the predicted curves. For these data, a range function is accounting quite well for the small but systematic *convergent* interactions found in these data.

EXAMPLE OF A FINAL SYSTEM STRUCTURE

The fit of the STFs to the data suggests the structure of the system within each experimental unit. Once an STF has been determined, subjective values associated with the system factors are also known.

The structure shown in Fig. 18 is the final structure that resulted from the hypotheses (two of which are depicted in Figs. 1 and 3) entertained by Veit, Callero, and Rose (1982). The difference between Fig. 18 and the hypothesized structure shown in Fig. 3 is that the intermediary factor "Execution Status" has been omitted.

The function selected as the appropriate STF is named at each path. At the top, a range function with a negative ω term best accounted for the immediate targeting experts' data. The negative ω indicates that divergent interactions were found among the six factors in this experimental unit. The interpretation of this finding is that the better the capability on one factor, the more of a difference it makes how good their capabilities are on the other factors. A range model with a positive ω term best accounted for the targeteers' data at the target identification path. The convergent interaction found here indicates that the more the targeteers knew about enemy emitters, the *less* of a difference it made how much information they had on enemy vehicles (and vice versa). (Better capabilities, however, always received a higher judgment.) The two hypothesized intermediary factors, Vehicles and Emitters, shown in Fig. 3, were retained. At the vehicle path, the range function with a negative ω term best accounted for overall divergent interactions found among the three factors at this experimental unit. The relative weight function with an initial impression (the third function shown in Table 2A) best accounted for the emitter data; the overriding trend in these data was independence among the factors on the target identification judgments. We will use this final system structure to illustrate the comparison of different systems defined by the same structure.

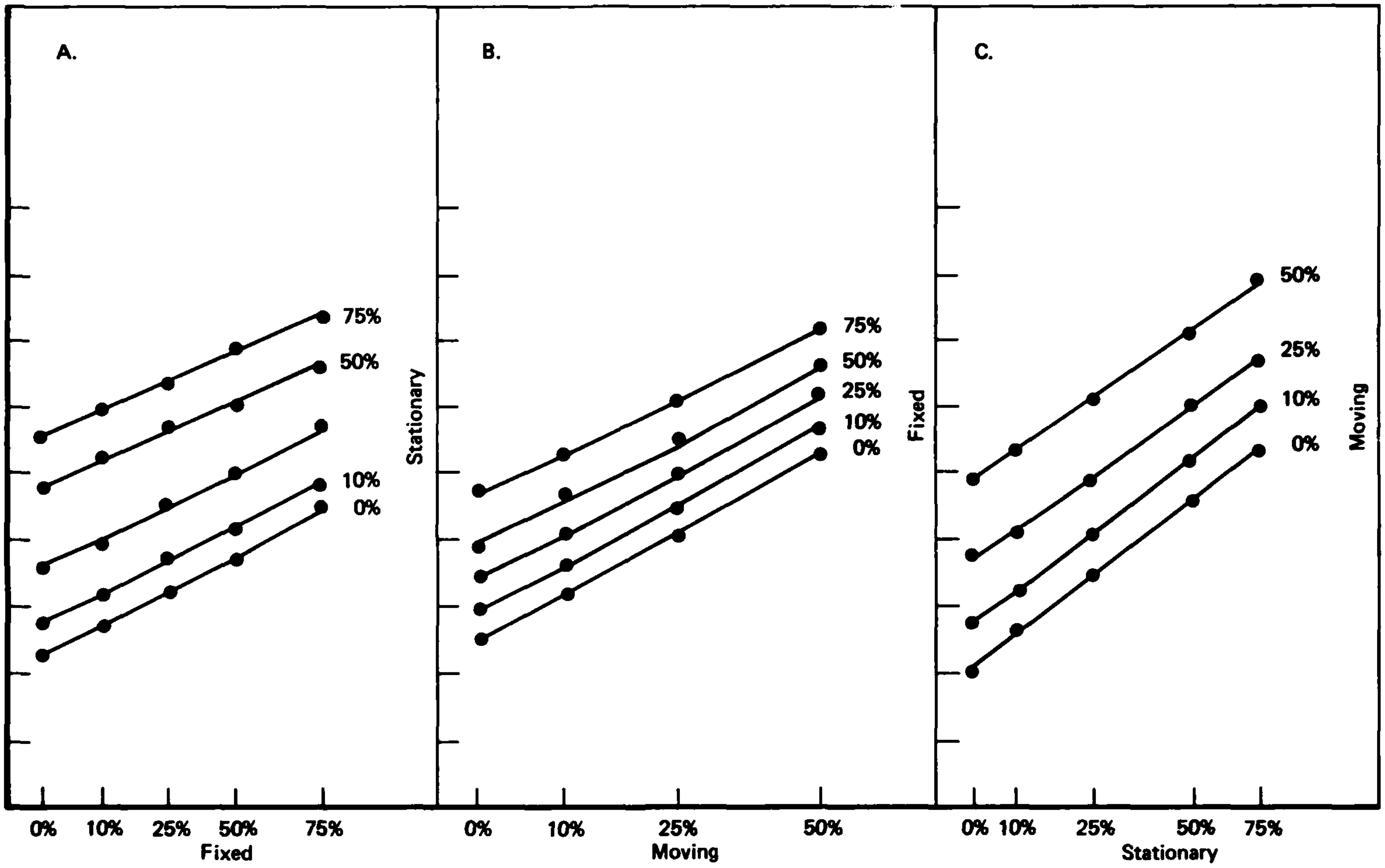


Fig. 17—Graphical assessment of the range model (in each panel, the curves represent the predicted values, the points represent the data)

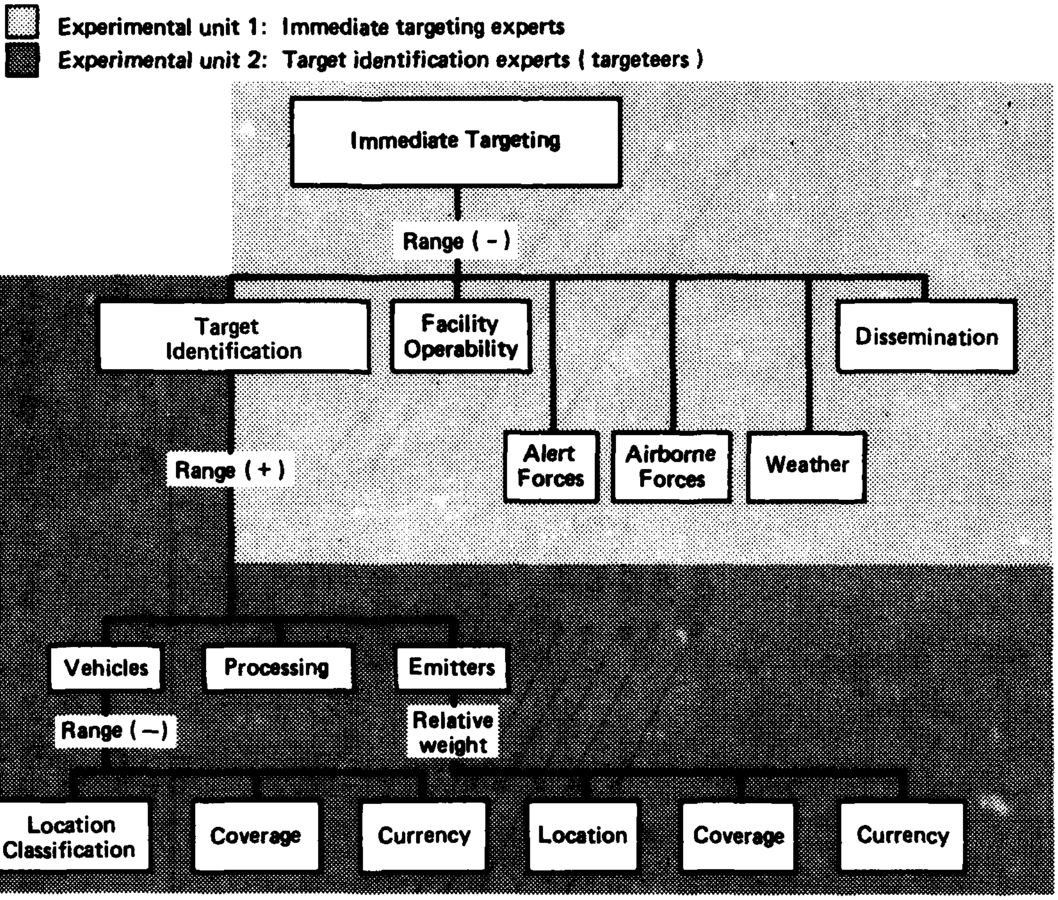


Fig. 18—Final structure and STFs corresponding to Fig. 1
(Veit, Callero, and Rose 1982)

V. EVALUATING SYSTEMS

Once the system's structure and STFs have been determined, it is possible to evaluate systems that differ in their capability levels. Capabilities are defined in terms of the system's primitive factor levels; systems can be considered *different* when they differ in at least one of these levels. The primitive factors in the structure shown in Fig. 18 are the location/classification, coverage, and currency factors associated with Vehicles; the location, coverage, and currency factors associated with Enemy Emitters; Processing; Facility Operability; Alert Forces; Airborne Forces; Weather; and Dissemination.

Determining the outcomes and suboutcomes of a system requires computing the STFs at the primitive factor paths, transferring these computed outputs into the STFs to which they are linked, and so forth until all STFs have been computed and all suboutcomes and the final outcome(s) have been obtained.

Subjective input values to the STFs at the primitive factor paths are needed to begin this process. Recall that when the program STEPIT tests the STF, it prints out all of the parameter values (e.g., factor weights, stimulus scale values, and the value for ω would be printed out for the range function). However, a scale value would not be available for a primitive factor level (stimulus) that was *not* manipulated in the experiment. This is not a problem when primitive factors are defined along a physical continuum (e.g., time, percent, distance), because it is possible to obtain the functional form of H in Eq. (2) by plotting subjective values associated with the manipulated factor levels as a function of the physical values. Thus, factor levels within the range manipulated can be obtained from the curve connecting those points. However, when factors are written descriptions, this plot results in a set of points that cannot be connected. If a system is defined by a different description, it is necessary to refield that experimental unit to obtain a scale value for the new primitive factor level.

Figure 19 exemplifies this evaluation process (see Table 1 for unabbreviated factor levels). Suppose it was of interest to compare the systems shown in Figs. 19A-C on how well the immediate targeting people thought they could do their job (the percent force application opportunities they thought they could exploit). The different systems are defined by the circled primitive factor levels.

First, it is necessary to obtain the subjective values associated with these factor levels. Figure 20 presents psychophysical functions for the vehicle coverage and currency factors and subjective points for the location/classification factor. The projected subjective values are the ones needed for the range function (shown at the top of Fig. 20) to compute the vehicle intermediary factor. Scale values for primitive factor levels at other paths would be obtained in the same way. This begins the computation procedure that continues to the top of the structure.

For the system shown in Fig. 19A, the STFs predict that the targeteers (lower portion of the structure) would perceive that they could identify about 33 percent of the important targets. Using this as the target identification factor level in the upper portion of the structure, immediate targeting experts perceive they could exploit about 48 percent of the important immediate targeting opportunities. Figure 19B shows that by increasing the targeteers' ability to identify targets to 68 percent, the ability to do immediate targeting increases to 52 percent,

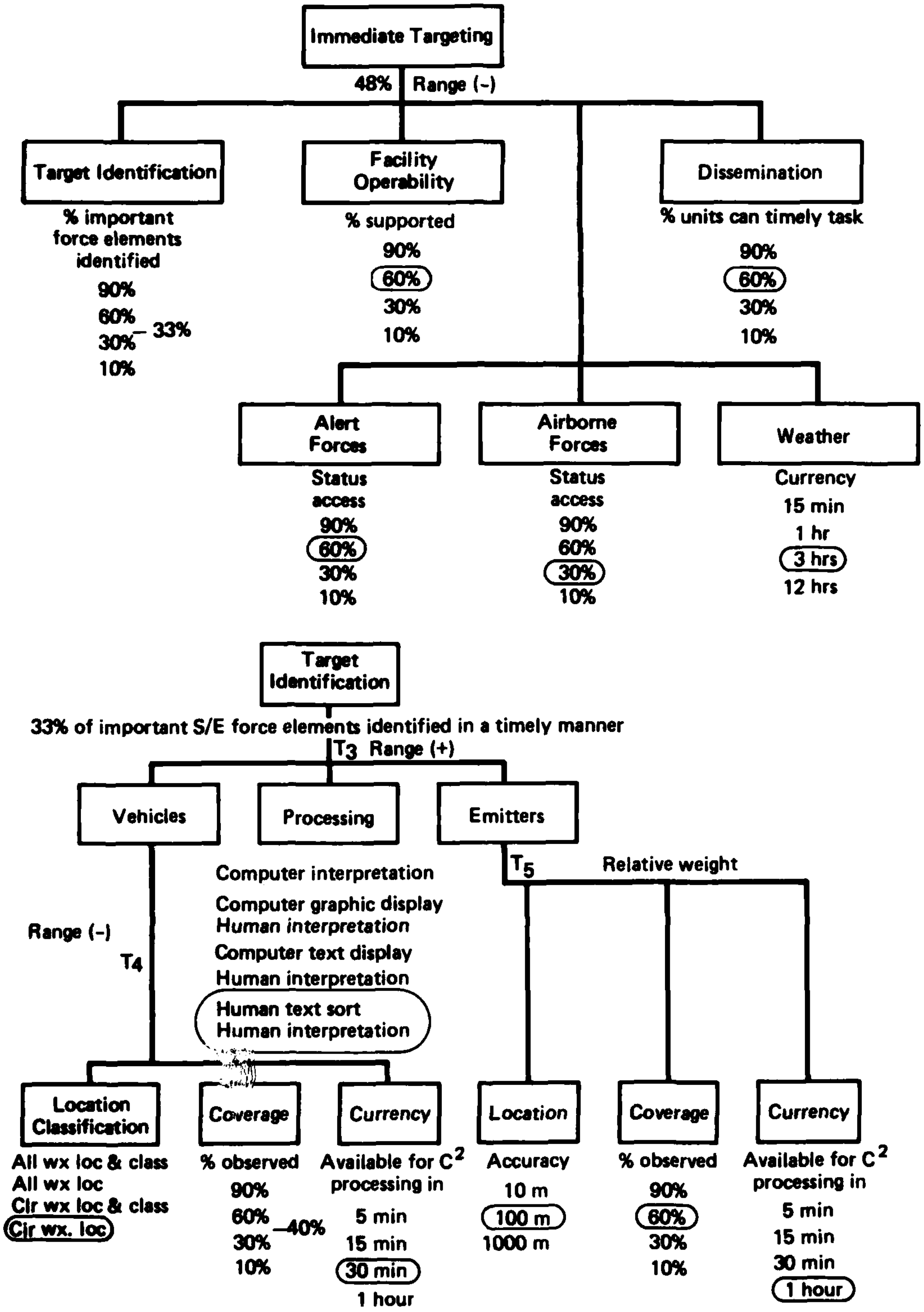


Fig. 19A—System comparisons

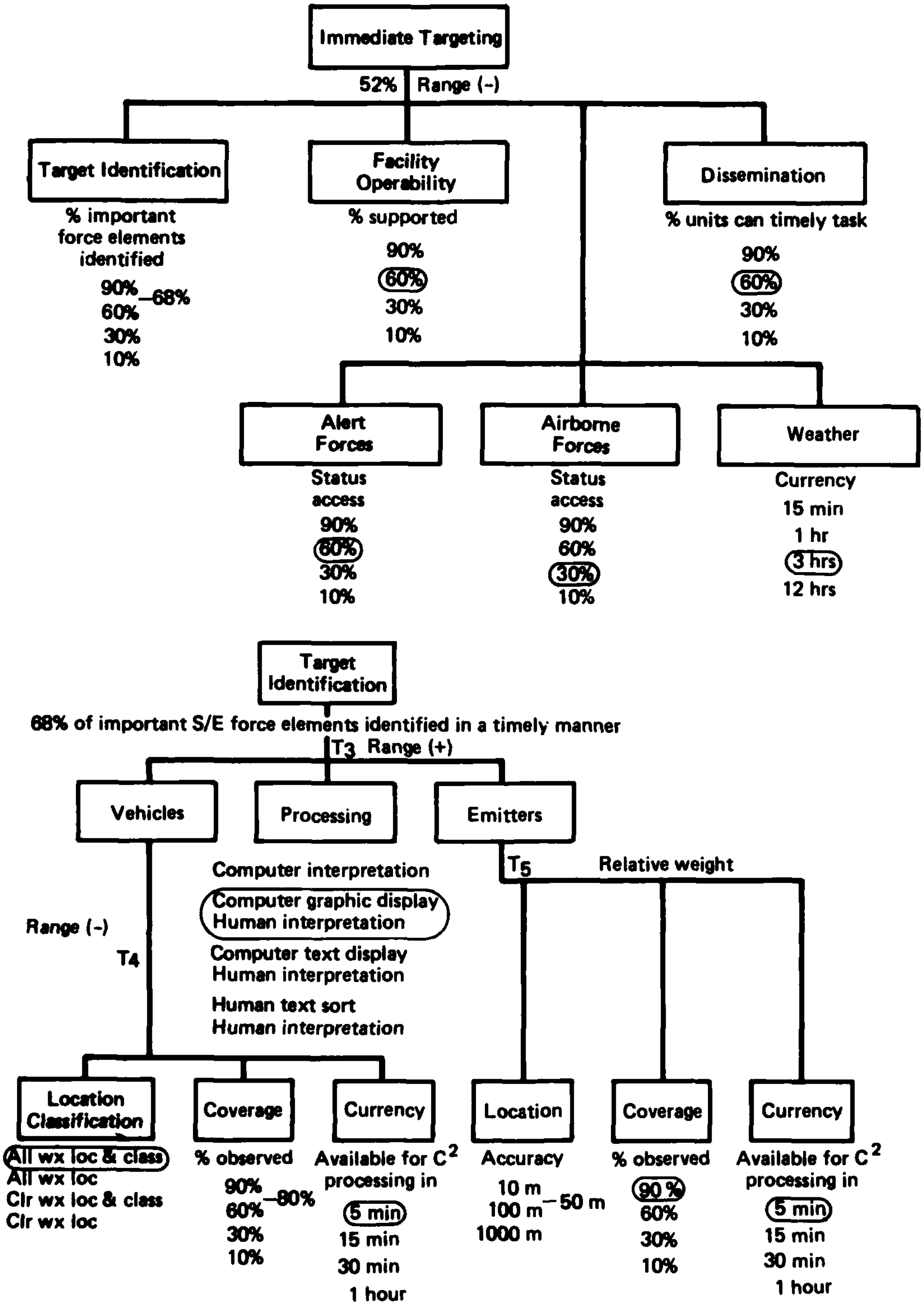


Fig. 19B—System comparisons

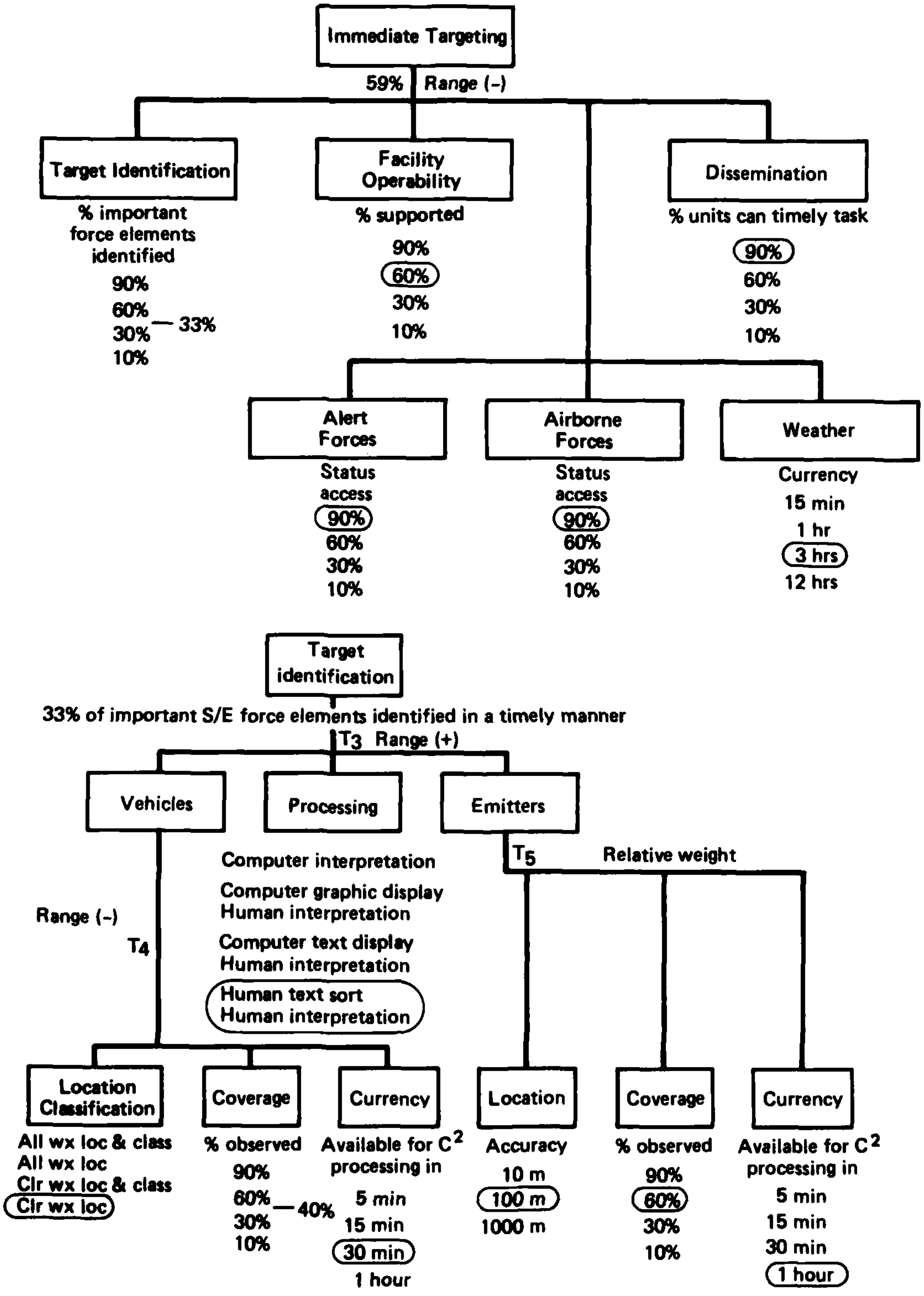


Fig. 19C—System comparisons

$$r = \frac{w_0 s_0 + w_L s_L + w_C s_C + w_T s_T}{w_0 + w_L + w_C + w_T} + \omega \, (s_{max} - s_{min})$$

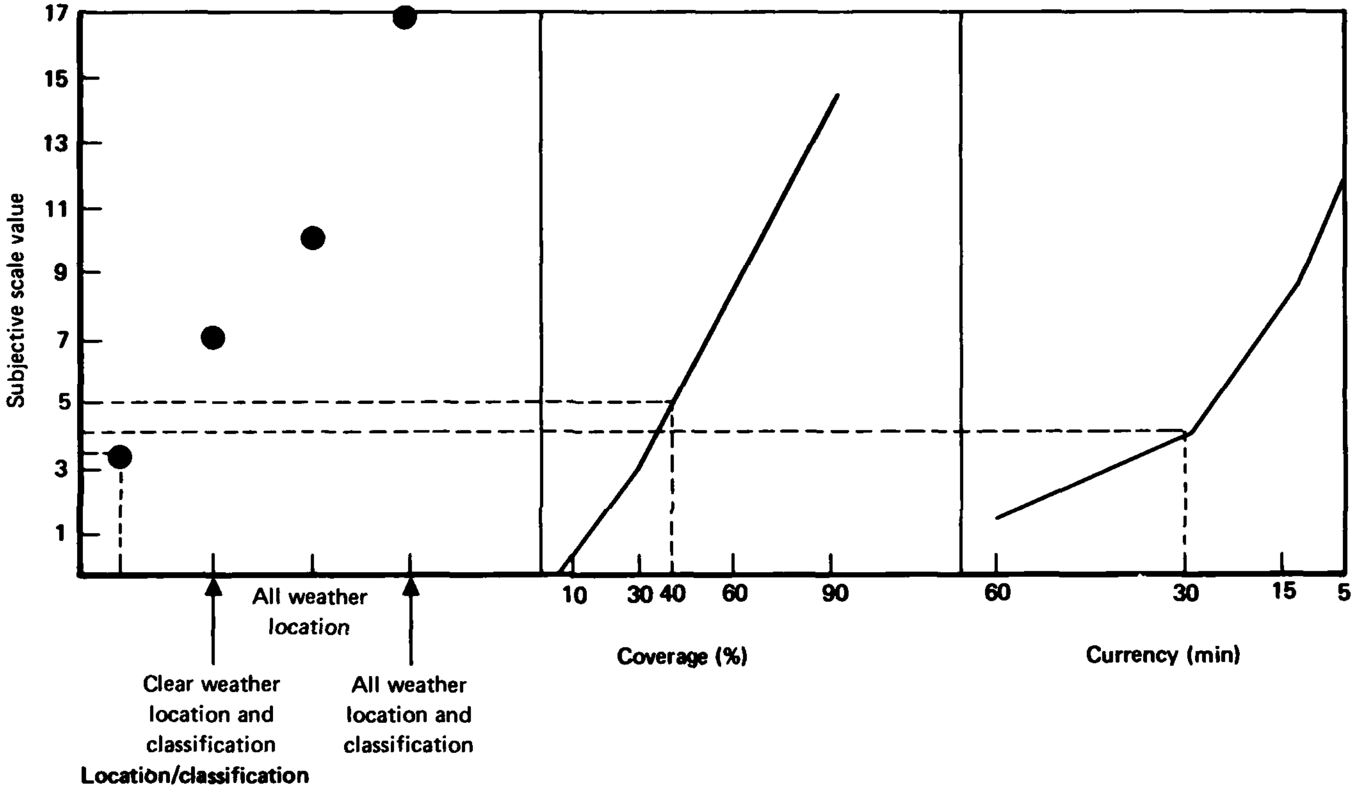


Fig. 20—Subjective scale values for the location/classification, coverage, and currency factors shown in Fig. 1

keeping the other immediate targeting capabilities at the levels shown in in Fig. 19.[1] In Fig. 19C, the system's capabilities are at their earlier levels (Fig. 19A), but the immediate targeting capabilities have been greatly improved in several areas (Alert Forces, Airborne Forces, and Dissemination),[2] increasing the ability to exploit immediate targeting opportunities to 59 percent.

The actual primitive factor levels selected in the evaluation would be determined from such things as systems that were being entertained for purchase, production, development, or present capability levels.

One can assess tradeoffs in the contribution of two factors to an outcome by examining a graph of STF predictions. Figure 21 shows this for the facility operability and dissemination factors depicted in Figs. 19A-C. For example, the subjective judgment on the y-axis is about the same for a dissemination level of 10 percent and a facility operability level of 90 percent as for a dissemination level of 60 percent and a facility operability level of 30 percent. One can assess other tradeoffs between these factors by drawing horizontal lines through the theoretical curves. For graphic tradeoffs among three factors, a graph for two factors such as that shown in Fig. 21 would be plotted at each level of the third factor. Four factors could be plotted by

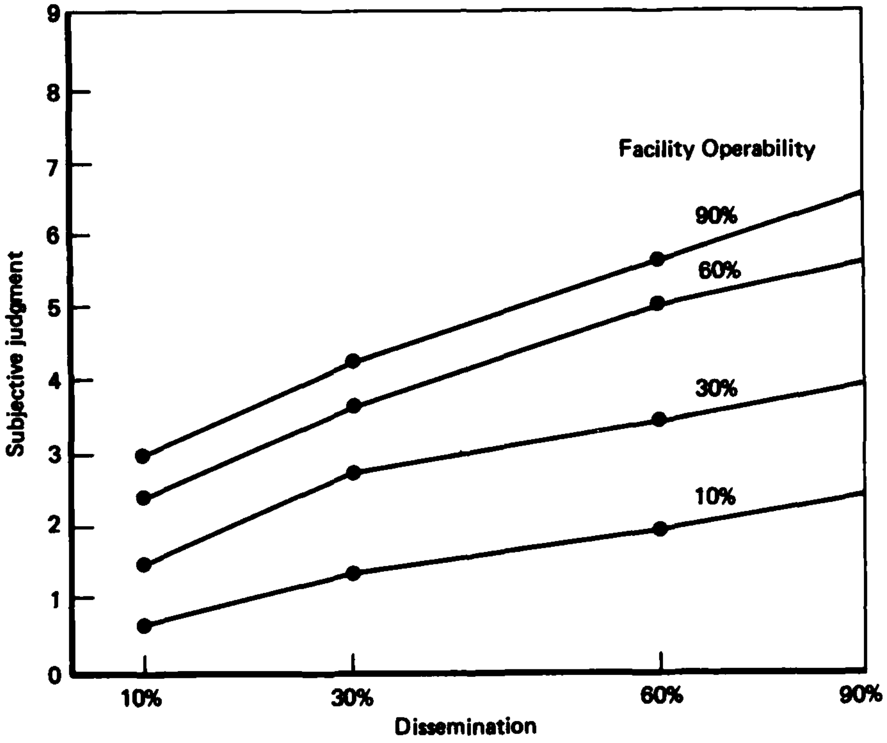


Fig. 21—Theoretical predictions

[1] An interpretation of these results is that the respondents felt they will be working in a target-rich environment (compared with available attack aircraft) and thus they put little value on identifying *more* important targets.

[2] These increased capabilities could result from the addition of better airborne capabilities.

graphing every factor level combination of two of the factors on the x-axis with a separate curve for the third factor, and a separate plot for each level of the fourth factor. Graphic procedures for evaluating tradeoffs would be especially useful in situations where the decision about *which* system changes to make involves only a few system factors.

An example of comparing two different systems defined by the larger structure in Fig. 7 is shown in Figs. 22 and 23. The two different systems are defined by the two different sets of primitive factor levels, delineated by rectangles in Figs. 22 and 23 (factor levels for the other primitive components have been held constant). The factors are the same for both the Plan and Control suboutcomes as follows:

> Precision: how accurate the information is that is reported to the command and control (C^2) system;
>
> Enemy Information Display: the means by which enemy information is displayed to the decisionmaker;
>
> Amount: the percent of enemy information data reported to the C^2 system;
>
> Currency: the frequency with which data are observed and reported to the C^2 system;
>
> Tactical Air: the percent of friendly tactical air systems about which information can be obtained in time for use by the decisionmakers;
>
> Friendly Information Display: the means by which friendly information is displayed in the C^2 system;
>
> Ground Force: the frequency that information on the status of friendly ground forces is collected and reported to the C^2 system.

The factor levels delineated for Plan and Control in Fig. 22 can be considered the baseline system. When STFs were computed up to the number of enemy targets neutralized in the three different target categories—Moving, Fixed, and Stationary—the results were the neutralization of 26 percent, 46 percent, and 31 percent of the important Moving, Fixed, and Stationary enemy targets. When these factors were changed to their top levels (with the exception of Enemy Information Display), it improved the capability of neutralizing important enemy targets very slightly—to 31 percent, 48 percent, and 37 percent for Moving, Fixed, and Stationary targets. Computing the STFs provides measures for each suboutcome in the structure so it is possible to see where changes are occurring. This information could be valuable to people who have to make decisions about what systems to develop or purchase. It provides them with systematic information about what difference changes make and where they occur in the system.

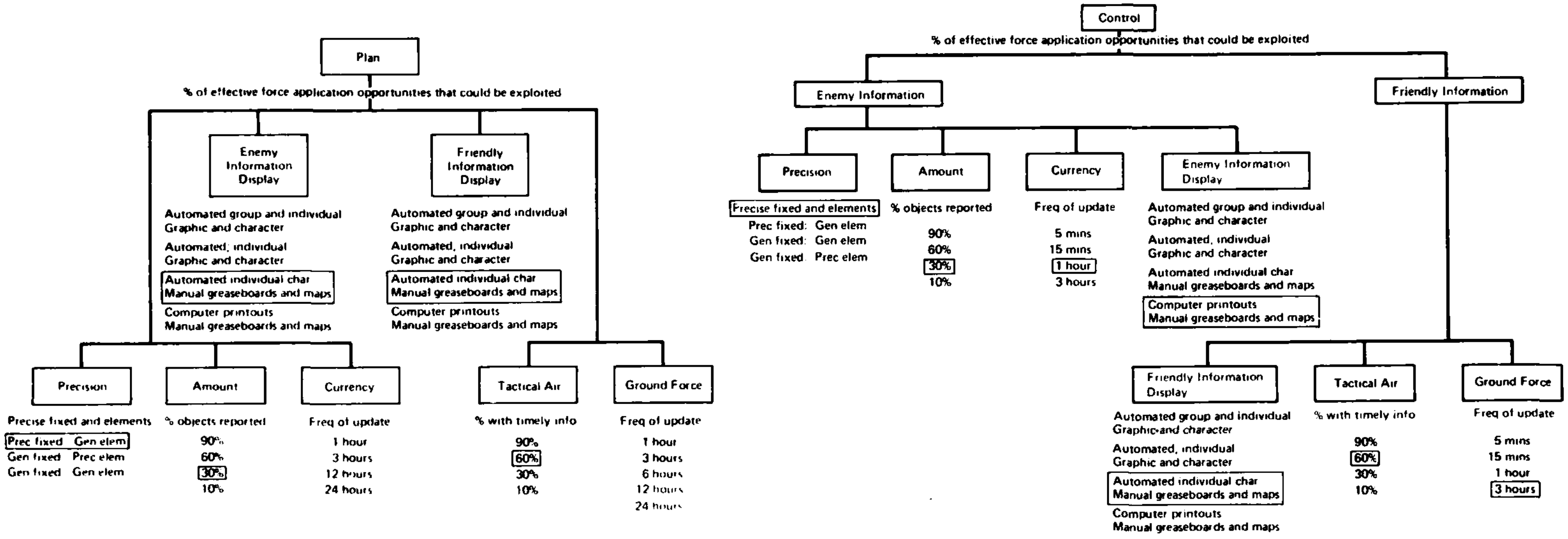


Fig. 22—One section of the command and control system shown in Fig.7
(primitive factor levels boxed)

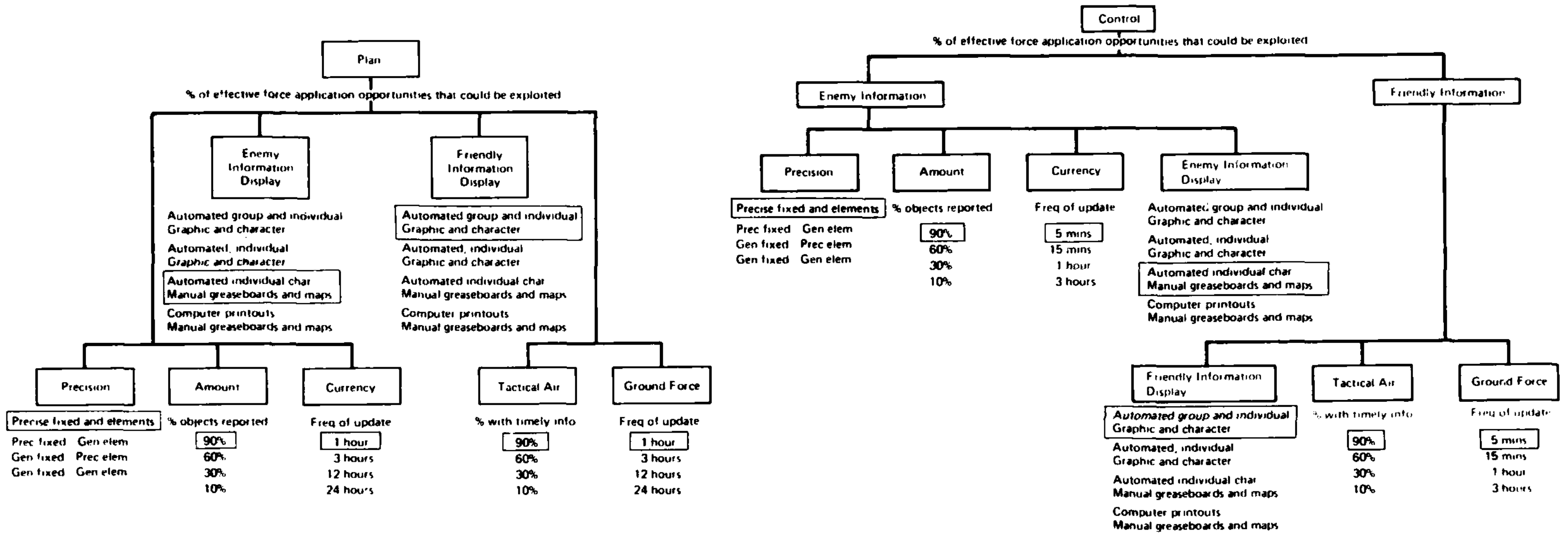


Fig. 23—Same section of Fig. 7 as shown in Fig. 22
(different primitive factor levels boxed)

VI. COMMENTS ON OTHER SUBJECTIVE APPROACHES TO ANALYZING SYSTEMS

USING "DIRECT" SCALES AND ASSUMED MODELS

The major advantage of the algebraic modeling approach is that it resolves the *testability* problems found in other commonly used approaches to measurement. For example, researchers using the multiple regression or multi-attribute utility theory function generally *assume* the validity of the function as the appropriate combination rule (T in Eq. (2)), as well as the validity of the weights and scale values used to compute the function. Thus, conclusions are based on untested premises. Further, the assumptions cannot in principle be tested given their experimental designs. Weights and scale values are generally obtained using "direct" scaling techniques (Stevens, 1946, 1957). In the direct approach, respondents assign numbers to stimulus descriptions according to instructions to obtain scales of "sensation." Many operational definitions have been proposed for obtaining direct scales (for example, Gardiner and Edwards (1975) recommend the category rating and magnitude estimation scales to obtain scales for the subjective expected utility model). However, psychologists have severely criticized the notion of operationally defining subjective scale values (Birnbaum, 1982; Birnbaum and Veit, 1974a; Krantz, 1972; Savage, 1966; Shepard, 1976; Treisman, 1964; Veit, 1978). Despite the lack of a testable basis for conclusions when models and scales are not obtained from a tested base, for many reasons (e.g., analyzing systems) people interested in obtaining subjective measures use and recommend these procedures (see, as examples, Gardiner and Edwards (1975); Keely, Andriole, and Daly, 1978; O'Conner, 1977; Martin, Bresnick, and Buede, 1981; Pirie, Frisvold, and Bresnick, 1981; and the PATTERN technique described in Waddington, 1977).

SAATY'S APPROACH

Saaty (1977) proposes another approach to obtain subjective measures fraught with measurement problems. The basic problems are that his ratio model is not adequately tested, and he provides no assessment of his aggregation model.

Saaty proposes that a ratio model underlies category ratings of ratios. This is a testable proposition using the factorial design of stimulus cues that he describes (given that the *respondent* and not the experimenter fills in the entire matrix). However, the appropriate test of the model is not the goodness-of-fit index he recommends. Indexes of fit can be high when deviations are significant and systematic. A way for both the researcher and the reader to assess the fit of a ratio model is to see graphs of the data. When responses are plotted as a function of the levels of one factor with a separate curve for each level of the other factor, the resulting curves should form a bilinear fan; *deviations* from the bilinear prediction of the model (which can be obtained from the analysis of variance) should be *nonsignificant* (Anderson, 1970; Birnbaum and Veit, 1974a; Veit, 1978). We have graphed the data presented in Saaty (1977) and have found them to be considerably different in form from the ratio model's predictions just described. If the model does not account for the data, what meaning can be attributed to the scale values derived from the model?

Further, suppose a ratio model did account for overt responses—that is, deviations from bilinearity were nonsignificant. There is a body of literature (Birnbaum, 1978, 1980; Birnbaum and Mellers, 1978; Veit, 1978, 1980) that suggests respondents take differences when instructed to take ratios and then take an exponential transformation of the differences. That is, when respondents are instructed to make ratio judgments, the combination function (T in Eq. (2)) is a subtractive model and J in Eq. (3) is an exponential transformation. Thus, the stimulus scale values would have to be obtained from the subtractive model after a logarithmic transformation were performed on the "ratios." The scale values derived from the subtractive model would be unique to an interval scale because any linear transformation of the scale values used in the subtractive model would reproduce the rank order of the data points in a factorial matrix.

It is interesting to note that scale values derived from a "correct" ratio model are unique only to a power transformation (Krantz and Tversky, 1971). Thus, even if the more stringent designs were needed to test whether T in Eq. (2) was subtractive or ratio (see for example, Birnbaum, 1980; Veit, 1978) *and* the data supported a ratio model, the scale values derived from the ratio model would be only a power transformation of the "true" values under the model; they would *not* be ratio scales as Saaty suggests. Therefore, it would be inappropriate to make ratio comparisons of the "weights" (scales obtained from the "ratio" model) within hierarchical levels (see for example, Alexander and Saaty, 1977).

SUMMARY REMARKS

Conclusions about what affects system outcomes must be credible. Their credibility rests on the measurement procedure used to produce them. When the procedure has entertained, rigidly tested, and rejected alternative hypotheses that would have led to *different* conclusions, then the hypothesis that is retained as the appropriate picture of the system is more believable.

The STF approach provides a subjective measurement framework for capturing the experts' perceptions of how his system functions. These perceptions provide information about how changes in system inputs affect outcomes. In this approach, scale values associated with system capabilities are theoretical parameters of a rigorously tested function. If a data array does not follow the configuration predicted by a hypothesized STF, both the STF and its parameters are rejected. The set of STFs and the structure that are retained as an explanation of the system emerge after stringent tests of alternative considerations.

BIBLIOGRAPHY

Alexander, J. M., and T. L. Saaty, "The Forward and Backward Processes of Conflict Analysis," *Behavioral Science,* 1977, 22.

Anderson, N. H., "Functional Measurement and Psychophysical Judgment," *Psychological Review,* 77, 1970, 153–170.

——, "Algebraic Rules and Psychological Measurement," *American Scientist,* 21, 1979, 201–215.

——, *Foundations of Information Integration Theory,* Academic Press, New York, 1981.

Birnbaum, M. H., "The Nonadditivity of Personality Impressions," *Journal of Experimental Psychology,* 102, 1974, 543–561 (monograph).

——, "Differences and Ratios in Psychological Measurement," in N. J. Castellan and Restle (eds.), *Cognitive Theory* (Vol. 3), Erlbaum, Hillsday, New Jersey, 1978.

——, "Comparison of Two Theories of 'Ratio' and 'Difference' Judgments," *Journal of Experimental Psychology: General.* 109, 1980, 304–319.

——, "Reason to Avoid Triangular Designs in Nonmetric Scaling," *Perception and Psychophysics,* 29, 3, 1981, 291–293.

——, "Controversies in Psychological Measurement," in B. Wegener (ed.), *Social Attitudes and Psychophysical Measurement,* Erlbaum, Hillsdale, New Jersey, 1982.

Birnbaum, M. H., and B. A. Mellers, "Measurement and the Mental Map," *Perception and Psychophysics,* 23, 1978, 403–408.

Birnbaum, M. H. and S. E. Stegner, "Source Credibility in Social Judgment: Bias, Expertise, and the Judge's Point of View," *Journal of Personality and Social Psychology,* 37, 1979, 48–74.

——, "Measuring the Importance of Cues in Judgment for Individuals: Subjective Theories of IQ as a Function of Heredity and Environment," *Journal of Experimental Social Psychology,* 17, 1981, 159–182.

Birnbaum, M. H., and C. T. Veit, "Scale Convergence as a Criterion for Rescaling: Information Integration with Difference, Ratio, and Averaging Tasks," *Perception and Psychophysics,* 16, 1974a, 276–282.

——, "Scale-free Tests of an Additive Model for the Size-weight Illusion," *Perception and Psychophysics,* 16, 1974b, 276–282.

Callero, M., B. J. Rose, and C. T. Veit, *Subjective Measurement of Tactical Air Command and Control: The STF Approach,* The Rand Corporation, 1984.

Chandler, J. P., "STEPIT—Finds Local Minima of a Smooth Function of Several Parameters," *Behavioral Science,* 14, 1969, 81–82.

Gardiner, P. C., and W. Edwards, "Public Values: Multiattribute-Utility Measurement for Social Decision Making," in M. F. Kaplan and S. Schwartz (eds.), *Human Judgment and Decision Processes,* Academic Press, New York, 1975.

Keely, C. W., S. J. Andriole, and J. A. Daly, "Computer-based Bayesian Information Processing," *Proceedings of the International Conference on Cybernetics and Society,* Part II, 1978.

Krantz, D. H., "Measurement Structure and Psychological Laws," *Science,* 175, 1972, 1427–1435.

Krantz, D. H., R. D. Luce, P. Suppes, and A. Tversky, *Foundations of Measurement,* Academic Press, New York, 1971.

Krantz, D. H., and A. Tversky, "Conjoint-measurement Analysis of Composition Rules in Psychology," *Psychological Review*, 78, 1971, 151–169.

Martin, A. W., T. A. Bresnick, and D. M. Buede, *Evaluation of Command and Control Centers*, Technical Report TR 81-3-328, Decisions and Designs, Inc., McLean, Va., 1981.

Miller, G. A., "The Magical Number Seven, Plus or Minus Two; Some Limits on Our Capacity for Processing Information," *The Psychological Review*, 63, 1956, 81–97.

Norman, K. L., "A Solution for Weights and Scale Values in Functional Measurement," *Psychological Review*, 83, 1, 1976, 80–84.

O'Conner, M. F., *Procedures for Assessing the Value of Command and Control Capabilities*, Technical Report 77-4, Decisions and Designs, Inc., McLean, Va., 1977.

Pirie, R. B., G. A. Frisvold, and T. A. Bresnick, *Application of Advanced Decision-Analytic Technology to Rapid Deployment Joint Task Force Problems*, Technical Report TR 81-7-328.13, Decisions and Designs, Inc., McLean, Va., 1981.

Rose, B. J., *The Relative Contributions of Physical, Mental, and Social Components to an Overall Measure of Health*, Doctoral dissertation, University of California, Los Angeles, 1980.

Rose, B. J., and M. Birnbaum, "Judgments of Differences and Ratios of Numerals," *Perception and Psychophysics*, 18, 1975, 94–200.

Saaty, T. L., "A Scaling Method for Priorities in Hierarchical Structures," *Journal of Mathematical Psychology*, 15, 1977, 234–281.

Savage, C. W., "Introspectionist and Behaviorist Interpretations of Ratio Scales of Perceptual Magnitudes," *Psychological Monograph*, 80, 1966, 1–32.

Shepard, R. N., "On the Status of 'Direct' Psychological Measurement," in C. W. Savage (ed.), *Minnesota Studies in the Philosophy of Science*, Vol. IX, University of Minnesota Press, Minneapolis, 1976.

Stevens, S. S., "On the Theory of Scales and Measurement," *Science*, 103, 1946, 677–680.

——, "On the Psychophysical Law," *Psychological Review*, 64, 1957, 153–181.

Treisman, M., "Sensory Scaling and the Psychophysical Law," *Quarterly Journal of Experimental Psychology*, 16, 1964, 11–22.

Veit, C. T., "Analyzing 'Ratio' and 'Difference' Judgments: A Reply to Rule and Curtis," *Journal of Experimental Psychology: General*, 109, 1980, 301–303.

——, "Ratio and Subtractive Processes in Psychophysical Judgment," *Journal of Experimental Psychology*, 107, 1978, 1, 81–107.

Veit, C. T., M. Callero, and B. J. Rose, "Demonstration of the Subjective Transfer Function Approach Applied to Air-Force-Wide Mission Area Analysis," The Rand Corporation, N-1831-AF, February 1982.

Veit, C. T., B. J. Rose, and J. E. Ware, "Effects of Physical and Mental Health on Health-state Preferences," *Medical Care*, 20, 4, 1982, 386–401.

Waddington, C. H., *Tools for Thought*, Basic Books, Inc., New York, 1977.